

Co-modularity and Co-community Detection in Large Networks

Abstract

This paper introduces the notion of co-modularity, to co-cluster observations of bipartite networks into co-communities. The task of co-clustering is to group together nodes of one type, whose interaction with nodes of another type are the most similar. The novel measure of co-modularity is introduced to assess the strength of co-communities, as well as to arrange the representation of nodes and clusters for visualisation. The existing non-parametric understanding of co-clustering is generalised in this paper, by introducing an anisotropic graphon class for realisations of bipartite networks. By modelling the smoothness of the anisotropic graphon directly, we can obtain a quantitative measure to determine the number of groups to be used when fitting co-communities, subsequently using the co-modularity measure to do so. We illustrate the power of our proposed methodology on simulated data, as well as examples from genomics and consumer-product reviews.

1 Introduction

Networks are used to parsimoniously represent relationships between entities of the same type. Classical analysis methods use parametric models of network data, such as degree-based and/or community-based models [Holland et al., 1983, Bickel and Chen, 2009, Rohe et al., 2011, Qin and Rohe, 2013, Wilson et al., 2013]. The last few years have seen a flurry of activity in statistical network analysis, see for example [Cai and Li, 2014, Amini et al., 2013, Amini and Levina, 2014]. One of the best-studied tools is the stochastic blockmodel [Bickel and Chen, 2009], and various extensions to it, see [Zhao et al., 2012, E. M. Airoldi, 2008, Gopalan and Blei, 2013]. The stochastic blockmodel is an important analysis tool, as it both groups variables of the same type, as well as simplifies the representation of their interactions. Various methods of fitting this model have been proposed, where maximizing modularity remains an important practical approach [Girvan and Newman, 2002, Bickel and Chen, 2009].

Recent work in clustering network nodes has generalised the applicability of the stochastic blockmodel, by showing that arbitrary exchangeable networks can be represented using a blockmodel [Diaconis, 1977, Bickel and Chen, 2009, Olhede and Wolfe, 2014]; such a representation is called a ‘network histogram’. This extends the applicability of the blockmodel, and naturally a degree-based correction as in [Zhao et al., 2012] can be added, to increase the parsimony of the representation. The network histogram, and the blockmodel in general, are piecewise-constant approximations of an underlying function, called the ‘graphon’ [Wolfe and Olhede, 2013]. The graphon function can be thought of as the generative mechanism of the data. The network histogram also provides a method to estimate the optimal number of blocks, or communities, which a valid blockmodel representation of the network comprises, if there is a smooth function in the graphon equivalence class. This is important and useful, because it means that the blockmodel can be used to identify, for example, an unknown number of communities in a social network, or an unknown number of functional subnetwork modules in a biological network. The network histogram method [Olhede and Wolfe, 2014] can be used to estimate the optimal granularity at which communities, or functional subnetwork modules, can be identified and isolated in social and biological networks, by fitting the stochastic blockmodel.

Studying relationships between variables of the same type is naturally of great utility; its simplest generalisation is to study relationships between variables of a different type; this is known as the co-clustering problem [Flynn and Perry, 2012, Choi et al., 2014, Madeira and Oliveira, 2004]. This problem can also be approached non-parametrically, as is made clear in [Choi et al., 2014]. To achieve consistent estimation, assumptions have to be made regarding the properties of the graphon function, where smoothness is standard [Olhede and Wolfe, 2014] and stronger assumptions [Airoldi et al., 2013] yield better estimation procedures when the stronger assumptions are justified. Assumptions made for the symmetric graphon function in the clustering problem need extension to the asymmetric graphon lying

behind the biclustering problem [Aldous, 1985]. To enable understanding of nonparametric estimation, we introduce the model of an anisotropic graphon, called the anisotropic graphon model.

Having provided a model for our set of relationships between two types of variables, we now need to infer them. We shall start from the modularity approach to recognising communities [Girvan and Newman, 2002], realising that extending such understanding to variables of different types is nontrivial [Aldous, 1985, Madeira and Oliveira, 2004]. Having recognised communities in both types of variables, we need to transform the clustering or grouping of both types of variables, into an ordering of groups. This is not inherent to the formulation of the Aldous-Hoover representation of the generating mechanism of the random array we are modelling, but is important for visualisation purposes. We also use the modularity to make this choice of visualisation.

To be able to use the modularity, we need to decide how many groups we are using in both variable types. This will be based on the model of the anisotropic graphon, and a choice of smoothness for the graphon function. We extend the work of [Olhede and Wolfe, 2014] to select the number of groups, adjusted for the anisotropic graphon model. All parameter choices are determined from the data, and a fully specified method of group allocation is given.

Finally, to demonstrate the power of our newly proposed method, we analyse two characteristic network data sets, namely a genomics data as well as a movie review data set. These two archetypical forms of data show the power of our proposed analysis methods, and enable us to discover both known and hitherto unknown characteristics of these two data sets.

This paper is organised as follows: Section 2 defines the stochastic block model, and gives the representation of an arbitrary separately exchangeable array. It also defines the co-modularity, and explains how the array data will be analysed. Section 3 describes how to choose the number of co-communities, and Section 4 shows how to determine them from data. Section 5 gives examples to illustrate the performance of our proposed method, and the supplement provides all proofs of the paper.

2 Co-modularity and Co-Community Detection

We begin this section by defining the degree-corrected stochastic co-blockmodel [Rohe and Yu, 2012, Flynn and Perry, 2012, Choi et al., 2014] together with notation; we then define a generalisation of this model based on the notion of the graphon. Following these model definitions, we give a definition of the Newman-Girvan modularity, and by analogy, we define a quantity which we term the ‘co-modularity’, and we specify an algorithm for maximising this quantity. We then show that under certain conditions, maximising the co-modularity in this way is equivalent to maximising the model likelihood of the specified degree corrected stochastic co-blockmodel.

Definition 1 (Degree-corrected stochastic co-blockmodel). *For $m, l \in \mathbb{N}^+$, define the set of X -nodes $\{1, \dots, m\}$, and the set of Y -nodes $\{1, \dots, l\}$. Denote an X -node grouping as $g_p^{(X)} \in G^{(X)}$, $p \in \{1, \dots, k^{(X)}\}$, and a Y -node grouping as $g_q^{(Y)} \in G^{(Y)}$, $q \in \{1, \dots, k^{(Y)}\}$, where $G^{(X)}$ and $G^{(Y)}$ are exhaustive lists of mutually exclusive X and Y -node groupings, respectively. Define map functions $z^{(X)}(i)$ and $z^{(Y)}(j)$, such that $g_p^{(X)} = \{i : z^{(X)}(i) = p\}$, and $g_q^{(Y)} = \{j : z^{(Y)}(j) = q\}$. Define co-community connectivity parameters $\theta \in [0, 1]^{k^{(X)} \times k^{(Y)}}$, where $\theta_{z^{(X)}(i), z^{(Y)}(j)}$ is the propensity of X -node i in group $z^{(X)}(i)$ to form a connection with Y -node j in group $z^{(Y)}(j)$. Define also node-specific connectivity parameters $\pi^{(X)} \in \mathbb{R}_{\geq 0}^m$ and $\pi^{(Y)} \in \mathbb{R}_{\geq 0}^l$. Let the elements of the adjacency matrix $\mathbf{A} \in \{0, 1\}^{m \times l}$ follow the law of:*

$$A_{ij} \sim \text{Bernoulli} \left(\pi_i^{(X)} \pi_j^{(Y)} \theta_{z^{(X)}(i), z^{(Y)}(j)} \right), 1 \leq i \leq m, 1 \leq j \leq l. \quad (1)$$

Then, A_{ij} is generated under the degree corrected stochastic co-blockmodel.

We note that the terminology ‘ X -nodes’ and ‘ Y -nodes’ is non-standard; we introduce it here, to increase clarity. To improve identifiability of parameters of the model in Definition 2, we introduce a specification favoured by many other authors [Newman, 2013], that $\theta_{z^{(X)}(i), z^{(Y)}(j)}$ may take only two values:

$$\theta_{p,q} = \begin{cases} \theta_{\text{in}}, & \text{if the pairing of } X\text{-node grouping } g_p^{(X)} \text{ with } Y\text{-node} \\ & \text{grouping } g_q^{(Y)} \text{ is a co-community,} \\ \theta_{\text{out}}, & \text{otherwise.} \end{cases} \quad (2)$$

We can also replace the Bernoulli model likelihood with a Poisson likelihood: because the Bernoulli success probability is typically small, and the number of potential edges (i.e., pairings of nodes) is large, a Poisson distribution with the same mean behaves very similarly, and so it makes little difference in practice [Zhao et al., 2012, Perry and Wolfe, 2012]. Its usage greatly simplifies the technical derivations. Hence, we calculate the model log-likelihood as follows (assuming $A_{ij} \in \{0, 1\}$ and therefore $A_{ij}! = 1$ for all i, j):

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\pi}^{(X)}, \boldsymbol{\pi}^{(Y)}; G^{(X)}, G^{(Y)}) \\ = \sum_{i=1}^m \sum_{j=1}^l A_{ij} \ln \left(\pi_i^{(X)} \pi_j^{(Y)} \theta_{z^{(X)}(i), z^{(Y)}(j)} \right) - \pi_i^{(X)} \pi_j^{(Y)} \theta_{z^{(X)}(i), z^{(Y)}(j)}. \end{aligned} \quad (3)$$

If we want to let the network grow, it would be impractical to fully specify more complicated versions of the parametric model of Definition 1, which completely account for all effects. Instead, we can make a non-parametric generalisation of this model incorporating more smoothing, based on the notion of the graphon. The graphon is a latent, smooth function which sets the probability between each pair of nodes, of a connection forming between that pair of nodes [Wolfe and Olhede, 2013]. In this setting, the graphon is not symmetric, due to the two different types of nodes modelled.

Definition 2. For the Lipschitz-continuous graphon $f \in L((0, 1)^2)$, with \mathbf{A} defined according to Definition 1, define connectivity functions $\phi^{(X)} \in L(0, 1)$ and $\phi^{(Y)} \in L(0, 1)$, and define latent orderings $\xi_i^{(X)} \stackrel{i.i.d.}{\sim} \mathcal{U}(0, 1)$ and (independently) $\xi_j^{(Y)} \stackrel{i.i.d.}{\sim} \mathcal{U}(0, 1)$ on the graphon margins of X and Y -nodes $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, l\}$ respectively. Then,

$$\mathbb{E}(A_{ij}) = f(\xi_i^{(X)}, \xi_j^{(Y)}) \cdot \phi^{(X)}(\xi_i^{(X)}) \cdot \phi^{(Y)}(\xi_j^{(Y)}). \quad (4)$$

The graphon f (Definition 2) can be considered an infinite-dimensional equivalent to $\theta_{p,q}$ (Definition 1), up to a re-ordering of the nodes defined by the orderings $\xi_i^{(X)}$ and $\xi_j^{(Y)}$ (which are always, to some extent, unidentifiable). The connectivity functions $\phi^{(X)}$ and $\phi^{(Y)}$ (Definition 2) are then similarly equivalent to the node-specific connectivity parameters $\boldsymbol{\pi}^{(X)}$ and $\boldsymbol{\pi}^{(Y)}$ (Definition 1). These functions $\phi^{(X)}$ and $\phi^{(Y)}$ model the general variability of connectivity strength throughout the network, whereas the graphon f models the tendency for regions of the network to aggregate into specific co-communities. The model of Definition 2 is a more general model which is specified similarly for any network size. However, as the networks we consider here are of fixed size, the degree corrected stochastic co-blockmodel (Definition 1) may be a more parsimonious choice. To estimate the generating mechanism of a bipartite network stably, Definition 2 must be replaced by a model with a limited number of parameters, i.e., Definition 1.

The Newman-Girvan modularity [Newman and Girvan, 2004] measures, for a particular partition of a network into communities, the observed number of edges between community members, compared to the expected number of edges between community members without the community partition. The Newman-Girvan modularity may be defined as follows:

Definition 3 (Newman-Girvan modularity). Define $\mathbf{A} \in \{0, 1\}^{n \times n}$ as a symmetric adjacency matrix representing a unipartite network with nodes $i \in \{1, \dots, n\}$, define \mathbf{d} as the degree vector of the nodes of this network, $d_i = \sum_{j=1}^n A_{ij}$, and define the normalising factor d^{++} as the total number of edges, $d^{++} = \sum_{i=1}^n d_i$. Define a community, or grouping, of nodes as $g \in G$, where G represents the set of all such groupings of nodes, define the map function $z(i)$ such that $g_a = \{i : z(i) = a\}$, and let $\mathbb{I}[z(i) = z(j)]$ specify whether nodes i and j appear together in any community g , such that:

$$\mathbb{I}[z(i) = z(j)] = \begin{cases} 1, & \text{if nodes } i \text{ and } j \text{ are grouped together} \\ & \text{in any community } g \in G, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the Newman-Girvan modularity Q_{NG} is defined as:

$$Q_{NG} = \frac{1}{d^{++}} \sum_{i=1}^n \sum_{j=1}^n \left[A_{ij} - \frac{d_i d_j}{d^{++}} \right] \cdot \mathbb{I}[z(i) = z(j)]. \quad (5)$$

The co-modularity is then defined by analogy with the Newman-Girvan modularity (Definition 3) as follows:

Definition 4 (Co-modularity). With \mathbf{A} given by Definition 1, define $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ as the degree vectors of the X and Y -nodes of the network, $d_i^{(X)} = \sum_{j=1}^l A_{ij}$ and $d_j^{(Y)} = \sum_{i=1}^m A_{ij}$, and define the normalising factor d^{++} as the total number of edges, $d^{++} = \sum_{i=1}^m d_i^{(X)} = \sum_{j=1}^l d_j^{(Y)}$. With $g^{(X)}$ and $g^{(Y)}$, $z^{(X)}$ and $z^{(Y)}$ also defined according Definition 1, let $c_t = \{p, q\} \in C$, $t = \{1, \dots, T\}$, if $T \neq 0$. The enumeration of the pair $\{p, q\}$ is arbitrary, and is to facilitate ease of access of the co-blocks in a chosen order. If $T = 0$, then by definition, $C = \emptyset$. The co-block c_t specifies that the X -node grouping $g_p^{(X)}$ is paired with the Y -node grouping $g_q^{(Y)}$; we refer to such a pairing as a ‘co-community’. Furthermore, let $\Psi(C; G^{(X)}, G^{(Y)}; i, j) \in \{0, 1\}$ specify whether nodes i and j appear together in any co-community $c \in C$, such that:

$$\Psi(C; G^{(X)}, G^{(Y)}; i, j) = \begin{cases} 1, & \text{if } \{z^{(X)}(i), z^{(Y)}(j)\} = c : c \in C, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the co-modularity Q_{XY} is defined as:

$$Q_{XY} = \frac{1}{d^{++}} \sum_{i=1}^m \sum_{j=1}^l \left[A_{ij} - \frac{d_i^{(X)} d_j^{(Y)}}{d^{++}} \right] \Psi(C; G^{(X)}, G^{(Y)}; i, j). \quad (6)$$

We note that for the co-modularity (unlike the Newman-Girvan modularity), we require a set of pairings of X -node groupings with Y -node groupings C , such that each $c_t \in C$ is a pairing of an X -node grouping $g_p^{(X)} \in G^{(X)}$ with a Y -node grouping $g_q^{(Y)} \in G^{(Y)}$. Also, due to the asymmetry of the co-clustering problem, $c_t = \{p, q\} \neq \{q, p\}$. This separately specified set of pairings C is not required in the case of the Newman-Girvan modularity, because in the unipartite network setting, there is only one type of node, and hence node groupings already ‘match-up’ with one another. This can be visualised, in the unipartite network setting, as community structure present along the leading diagonal of the adjacency matrix, if the nodes are ordered by community. In the co-community setting, an X -node grouping $g^{(X)}$ may be paired in C with many, with one, or with no Y -node groupings $g^{(Y)} \in G^{(Y)}$, and equivalently a Y -node grouping $g^{(Y)}$ may be paired in C with many, with one, or with no X -node groupings $g^{(X)} \in G^{(X)}$. Further, if the X -nodes and Y -nodes of the network are arranged in the adjacency matrix according to the groupings $g^{(X)}$ and $g^{(Y)}$, there is no reason co-communities should appear along the leading diagonal. Hence, the function Ψ in Equation 6 generalises the role of the indicator function in Equation 5. We also note that sometimes in practice, we must relax the requirement of $\Psi \in \{0, 1\}$; the reason for this becomes clear in the technical derivations in Supplement A which relate to Algorithm 1 (which follows next).

Community detection of k communities can be performed by fitting the degree-corrected stochastic blockmodel. This is equivalent, under many circumstances, to spectral clustering [Bickel and Chen, 2009, Riolo and Newman, 2012, Newman, 2013], which may be carried out by grouping the nodes into k clusters in the space of the eigenvectors corresponding to the 2nd to k^{th} greatest eigenvalues of the Laplacian $\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$, where \mathbf{D} is the diagonal matrix of the degree distribution. Co-community detection in a bipartite network of nodes attributed to the variables X and Y (respectively, X -nodes and Y -nodes), can equivalently be performed by degree-corrected spectral clustering [Dhillon, 2001].

A procedure to find an assignment of X and Y -nodes to $k^{(X)}$ X -node groupings (‘row clusters’) and $k^{(Y)}$ Y -node groupings (‘column clusters’) respectively, which finds a (possibly locally) optimum value of the co-modularity Q_{XY} , is specified in Algorithm 1:

Algorithm 1. With \mathbf{A} and Q_{XY} defined as in Definition 1, and $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ defined as in Definition 4:

1. Calculate the co-Laplacian \mathbf{L}_{XY} [Dhillon, 2001] as:

$$\mathbf{L}_{XY} = \left(\mathbf{D}^{(X)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(Y)} \right)^{-1/2}, \quad (7)$$

where $\mathbf{D}^{(X)}$ and $\mathbf{D}^{(Y)}$ are the diagonal matrices of $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$, respectively.

2. Calculate the singular value decomposition (SVD) of the co-Laplacian \mathbf{L}_{XY} .
3. Separately cluster the X and Y -nodes in the spaces of the left and right singular vectors corresponding to the 2^{nd} to $k^{(X)\text{th}}$ and 2^{nd} to $k^{(Y)\text{th}}$ greatest singular values, respectively, of this SVD of \mathbf{L}_{XY} .

Technical derivations relating to Algorithm 1 appear in Supplement A, and are based on arguments made previously in the context of unipartite (symmetric) community detection [Newman, 2013], extending them to this context of (asymmetric) co-community detection. We note in particular, that the notion of modularity assumes that within-community edges are more probable than between-community edges, and therefore modularity maximisation is only consistent if constraints are applied to ensure this assumption holds [Zhao et al., 2012]. In the community detection setting, under suitable constraints, the solutions which maximise model likelihood and modularity are identical [Bickel and Chen, 2009].

Proposition 1. *The solution which maximises the model likelihood specified in equation 3, subject also to the constraint of equation 2, is equivalent to the maximum co-modularity assignment obtained via Algorithm 1.*

Proof. The proof appears in Supplement B. It extends arguments made previously in relation to community detection [Newman, 2013] to this context of co-community detection. \square

3 Selecting the Number of Co-communities

In order to use Algorithm 1 to carry out co-community detection, we must specify the number of X -node groupings $k^{(X)}$, and the number of Y -node groupings $k^{(Y)}$. The network histogram method of fitting the stochastic blockmodel [Olhede and Wolfe, 2014] in the unipartite/symmetric community detection setting provides a rule-of-thumb method for selecting the optimal number of communities, or blocks, in the model. Fitted in this way, the blockmodel is a valid representation of a network, whatever the generating mechanism of that network, as long as this generating mechanism results in an exchangeable network. The network histogram approximates the graphon, which is a continuous function: the nodes correspond to discrete locations along the graphon margins, ordered in an optimal way to satisfy the smoothness requirement of the graphon. The graphon oracle [Wolfe and Olhede, 2013, Olhede and Wolfe, 2014] defines a good ordering of the nodes, according to graphon smoothness, and community structure. This information is not available in practice, but it can be used to bound the mean integrated squared error of the network histogram approximation to the graphon. This ordering naturally corresponds to community assignments, and the number of communities, or blocks, is determined by the smoothness of the graphon. An intuition for this is by analogy with a wave: if there are many peaks over a fixed distance (i.e., short wavelength), the maximum gradient of the wave will be large, whereas if there are few peaks over the same fixed distance (i.e., long wavelength), the maximum gradient will be small. Similarly, the more communities, or peaks, that there are in the graphon, the greater the maximum gradient of the graphon will be, and, correspondingly, the less smooth it will be.

3.1 Finding the optimal numbers of X and Y -node groupings

In this section we define the anisotropic graphon, which allows us to determine an optimal number of X and Y -node groupings, $k^{(X)}$ and $k^{(Y)}$, from which co-communities can be identified. This relates closely to the network histogram method in the symmetric unipartite community detection setting [Olhede and Wolfe, 2014]. In the unipartite community detection setting, the graphon is a symmetric limit object bounded on $(0, 1)^2$. It is symmetric because in that setting, the set of X -nodes is the same as the set of Y -nodes, and hence the smoothness is the same with respect to the corresponding orthogonal directions on the graphon. In contrast, in this co-community detection setting the graphon is asymmetric, having different smoothnesses with respect to the X and Y -nodes. Hence, we refer to this as the ‘anisotropic graphon’, which is similarly a limit object bounded on $(0, 1)^2$. To aid our analyses, we can stretch the anisotropic graphon so that it has the same smoothness with respect to the X -nodes, and with respect to the Y -nodes. It is easy to see that such a transformation exists for all anisotropic graphons. We refer to the result of stretching the anisotropic graphon in this way, as the ‘equi-smooth graphon’. Without loss of generality, this transformation can be expressed as a stretch of scale-factor γ with respect to the

X -nodes, and a simultaneous stretch of scale-factor $1/\gamma$ with respect to the Y -nodes. We refer to γ as the anisotropy factor. This is formalised as follows.

Definition 5. For the Lipschitz-continuous anisotropic graphon $f \in L((0,1)^2)$ defined according to Definition 2, let the anisotropy factor γ define the linear-stretch transformation which maps f onto the Lipschitz-continuous equi-smooth graphon $\tilde{f} \in L((0,\gamma) \times (0,1/\gamma))$. Then,

$$f(x, y) = \tilde{f}(\gamma x, y/\gamma). \quad (8)$$

Lipschitz-continuity, in this context, means that the smoothness of the graphon (anisotropic or equi-smooth) is upper-bounded, and we use this bound to calculate the optimal number of X and Y -node groupings.

To determine the optimal number of X and Y -node groupings, $k^{(X)}$ and $k^{(Y)}$, we set these $k^{(X)}$ and $k^{(Y)}$ so as to minimise the mean integrated squared error (MISE) of the blockmodel approximation of the graphon. Following a methodology which is closely related to the network histogram estimator in the symmetric (unipartite) community detection setting [Olhede and Wolfe, 2014], making use of the graphon oracle estimator, an upper bound can be calculated on this MISE, from a bias-variance decomposition, as follows:

Lemma 1. With \mathbf{A} , m , l , $g^{(X)} \in G^{(X)}$, and $g^{(Y)} \in G^{(Y)}$ defined according to Definition 1, let ρ be a deterministic scaling constant which specifies the expected number of edges in the network, such that:

$$\rho = \mathbb{E} \left(\frac{1}{ml} \sum_{j=1}^l \sum_{i=1}^m A_{ij} \right),$$

and define piecewise block-approximations to the adjacency matrix, for each pairing of a set of X -nodes $g^{(X)}$ with a set of Y -nodes $g^{(Y)}$, as:

$$\bar{A}_{p,q} = \frac{\sum_{i \in g_p^{(X)}, j \in g_q^{(Y)}} A_{ij}}{|g_p^{(X)}| |g_q^{(Y)}|}$$

where $|\cdot|$ represents cardinality. With $z^{(X)}$ and $z^{(Y)}(j)$ defined according to Definition 1, $\xi^{(X)}$ and $\xi^{(Y)}$ defined according to Definition 2, and f defined according to Definition 5, define alternative map functions $\tilde{z}^{(X)}(i')$, $i' \in \{1, \dots, m\}$, and $\tilde{z}^{(Y)}(j')$, $j' \in \{1, \dots, l\}$, which take the ordered locations of the X and Y -nodes respectively along the graphon margins, as specified by $\xi^{(X)}$ and $\xi^{(Y)}$, and return the corresponding X and Y -node groupings, such that $\tilde{z}^{(X)}(\lceil m \cdot \xi_i^{(X)} \rceil) = z^{(X)}(i)$, and $\tilde{z}^{(Y)}(\lceil l \cdot \xi_j^{(Y)} \rceil) = z^{(Y)}(j)$. Define the graphon oracle estimator as:

$$\hat{f}(x, y) = \hat{\rho}^{-1} \bar{A}_{\tilde{z}^{(X)}(\lceil lx \rceil), \tilde{z}^{(Y)}(\lceil my \rceil)}, \quad (9)$$

and let:

$$\iint_{(0,1)^2} f(x, y) dx dy = 1. \quad (10)$$

With \tilde{f} and γ defined as in Definition 5, let \tilde{M} be the maximum gradient of \tilde{f} , and let $h^{(X)}$ and $h^{(Y)}$ be ‘bandwidth’ model parameters with respect to the X and Y nodes respectively. Then, the graphon oracle upper bound on the MISE of the blockmodel estimate of the graphon function \hat{f} is:

$$\begin{aligned} \text{MISE}(\hat{f}) &\leq \tilde{M}^2 \left\{ \gamma^2 \cdot \frac{(h^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h^{(Y)})^2}{l^2} \right\} \\ &\quad + 2\tilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\} \{1 + o(1)\} + \frac{1}{\rho \cdot h^{(X)} \cdot h^{(Y)}} \{1 + o(1)\}. \end{aligned} \quad (11)$$

Proof. See supplement C. □

We note that the sets of nodes represented by the groupings $g^{(X)} \in G^{(X)}$ and $g^{(Y)} \in G^{(Y)}$ are contiguous along the graphon margins (corresponding to the canonical graphon ordering, [Airoldi et al. \[2013\]](#), [Chan and Airoldi \[2014\]](#)), but that these nodes are not contiguous along the adjacency matrix margins. Thus, we need to specify how nodes map to the groupings $g^{(X)}$ and $g^{(Y)}$ in a different way for the graphon, as compared to the adjacency matrix. This difference is accounted for by using different mapping functions: $\tilde{z}^{(X)}(i')$ and $\tilde{z}^{(Y)}(j')$ for the graphon, and $z^{(X)}(i)$ and $z^{(Y)}(j)$ for the adjacency matrix. I.e., $\tilde{z}^{(X)}(i')$ and $\tilde{z}^{(Y)}(j')$ are required to specify the (contiguous) ranges and locations of the X and Y -node groupings $g^{(X)}$ and $g^{(Y)}$ on the graphon margins, and equivalently $z^{(X)}(i)$ and $z^{(Y)}(j)$ for their (non-contiguous) locations on the adjacency matrix margins.

Using the MISE formulation of Lemma 1, we can estimate the optimal numbers of X and Y -node groupings, $k^{(X)}$ and $k^{(Y)}$.

Proposition 2. *With m and l defined as in Definition 1, and \widetilde{M} and ρ defined as in Lemma 1, the optimal number of X and Y -node groupings, $k^{(X)}$ and $k^{(Y)}$ respectively, are:*

$$k^{(X)} = \gamma \cdot (ml)^{\frac{1}{4}} \cdot \left(2\rho\widetilde{M}^2\right)^{\frac{1}{4}} \quad (12)$$

and

$$k^{(Y)} = \frac{1}{\gamma} \cdot (ml)^{\frac{1}{4}} \cdot \left(2\rho\widetilde{M}^2\right)^{\frac{1}{4}}. \quad (13)$$

Proof. The proof of this proposition is developed from the equivalent proof for the case of the isotropic graphon (corresponding to community detection in unipartite networks) [\[Olhede and Wolfe, 2014\]](#). The optimal bandwidths $h^{(X)*}$ and $h^{(Y)*}$ can be found by differentiating the expression for the MISE of equation 11 with respect to $h^{(X)}$ and setting to zero, and doing the same with respect to $h^{(Y)}$, and combining the resulting equations. To calculate $k^{(X)}$ and $k^{(Y)}$, substitute these optimal bandwidths $h^{(X)*}$ and $h^{(Y)*}$ into $k^{(X)} = m/h^{(X)*}$ and $k^{(Y)} = l/h^{(Y)*}$, which leads to equations 12 and 13. \square

We note that the above proof of Proposition 2 implies constant group sizes for the X -nodes, and constant group sizes for the Y -nodes. This assumption is relaxed in the practical implementation of this methodology we propose: this point is discussed further in Section 3.2.

3.2 Practical estimation of the number of X and Y -node groupings

We implement spectral clustering by including a standard k -means step, to group the X and Y -nodes in the spaces of the left and right singular vectors corresponding to the 2nd to $k^{(X)}$ th and 2nd to $k^{(Y)}$ th greatest singular values, respectively, of the singular value decomposition of the co-Laplacian \mathbf{L}_{XY} (equation 7). This k -means step does not produce identical group sizes, however we note that the estimates of $k^{(X)}$ and $k^{(Y)}$ defined according to equations 12 and 13 assume that the X and Y node groupings are the same size (i.e., that the blocks in the blockmodel are all the same size with respect to the X -nodes, and separately with respect to the Y -nodes). We relax this requirement in practice, because after examining several empirical data-sets of the type presented in the next section, we observed that the group sizes produced by this type of regularised degree-corrected spectral clustering, tend not to vary significantly in size (there are no ‘giant clusters’). Further, this requirement of identical group sizes is not physically realistic in the practical examples we present in the next section, and in many other real scenarios.

To estimate \widetilde{M} and γ , we approximate the maximum slope of the graphon separately in the directions corresponding to the X and Y -nodes, by considering the top component of the singular value decomposition of the adjacency matrix \mathbf{A} . This is equivalent to the rule-of-thumb procedure in the network histogram method, in the symmetric/unipartite community detection scenario [\[Olhede and Wolfe, 2014\]](#). The top left and right singular vectors are ordered, and their gradients and values at their midpoints (the expected points of maximum slope) are estimated as \hat{p}_X and \hat{b}_X respectively for the X -nodes, and \hat{p}_Y and \hat{b}_Y respectively for the Y -nodes. By thinking of this singular value decomposition as a factorisation of the scaled, discrete-sampled graphon (i.e., the ordered adjacency matrix), denoting the greatest singular value as ν , leads to the linear approximations for the maximum gradient of the isotropic graphon M in the directions of the X and Y -nodes, M_X and M_Y respectively:

$$\hat{M}_X = \frac{\nu}{\rho} \hat{p}_X \hat{b}_Y m, \quad \hat{M}_Y = \frac{\nu}{\rho} \hat{b}_X \hat{p}_Y l,$$

where m and l are the number of X and Y -nodes respectively (as previously defined). These factors m and l take account of the fact that the isotropic graphon margins are bounded on $[0, 1]$, whereas the adjacency matrix margins take the values $\{1, \dots, m\}$ and $\{1, \dots, l\}$, and the edge density factor ρ (defined as in Lemma 1) normalises with respect to the adjacency matrix realisation, such that the above estimates are independent of edge density ρ . The linear stretch transformation γ defines the maximum gradients of the equi-smooth graphon as $\hat{M}_X = \gamma \hat{M}_X$ and $\hat{M}_Y = \hat{M}_Y / \gamma$ respectively, and hence an estimate of the squared maximum gradient of the isotropic graphon can be found as:

$$\hat{\bar{M}}^2 = \gamma^2 \cdot \hat{M}_X^2 + \frac{1}{\gamma^2} \cdot \hat{M}_Y^2 = \frac{\nu^2}{\rho^2} \left(\gamma^2 \cdot \hat{p}_X^2 \hat{b}_Y^2 m^2 + \frac{1}{\gamma^2} \cdot \hat{b}_X^2 \hat{p}_Y^2 l^2 \right).$$

Using the assumption that the equi-smooth graphon is Lipschitz-continuous, with the same upper-bound on its smoothness with respect to both the X and Y nodes, i.e., $\hat{M}_X = \hat{M}_Y$, $\implies \gamma \hat{M}_X = \hat{M}_Y / \gamma$, we can estimate γ as:

$$\hat{\gamma}^2 = \frac{\hat{M}_Y}{\hat{M}_X}. \quad (14)$$

3.3 Model simplifications

We can draw an analogy between bandwidth estimation in the anisotropic graphon, and the anisotropic kernel [Wand and Jones, 1993, Duong and Hazelton, 2003]. Similar to bivariate kernel density estimation, we may be able to achieve a more parsimonious model, if we can justifiably assume that the smoothness of the anisotropic graphon is the same with respect to both the X and Y nodes. This is the same as saying that the anisotropy factor $\gamma \approx 1$, and that $\hat{M}_X \approx \hat{M}_Y$.

Proposition 3. *With γ defined as in Definition 5, testing the following null and alternative hypotheses:*

$$H_0 : \quad \gamma = 1, \quad H_1 : \quad \gamma \neq 1,$$

under the null, the estimated anisotropy constant $\hat{\gamma}$ follows the law of:

$$\gamma^2 \sim \mathcal{N}(1, \tau^2),$$

where the variance τ^2 is estimated from the linear model estimates of \hat{b}_X , \hat{p}_X , \hat{b}_Y and \hat{p}_Y , such that:

$$\begin{aligned} \hat{\tau}^2 = & \frac{\widehat{\text{Var}}(\hat{b}_X)}{\hat{b}_X} + \frac{\widehat{\text{Var}}(\hat{p}_Y)}{\hat{p}_Y} + \frac{\widehat{\text{Var}}(\hat{p}_X)}{\hat{p}_X} + \frac{\widehat{\text{Var}}(\hat{b}_Y)}{\hat{b}_Y} \\ & + 2 \frac{\widehat{\text{Cov}}(\hat{b}_X, \hat{p}_X)}{\hat{b}_X \hat{p}_X} + 2 \frac{\widehat{\text{Cov}}(\hat{b}_Y, \hat{p}_Y)}{\hat{b}_Y \hat{p}_Y}. \end{aligned}$$

Proof. See supplement C. □

If we fail to reject H_0 in Proposition 3, then we take it that $\hat{M}_X \approx \hat{M}_Y$, and $\gamma \approx 1$, and hence we have that $k^{(X)} = k^{(Y)} = (ml)^{\frac{1}{4}} \cdot (2\rho\hat{M}^2)^{\frac{1}{4}}$, i.e., there are the same number groupings of X -nodes and of Y -nodes. This assumption is implicitly made in widely-used previous solutions to the co-community detection problem, as in [Dhillon, 2001].

It is also worth noting that if the number of X and Y nodes, m and l respectively, are very different, then this does not preclude $\gamma \approx 1$: there are, in effect, two independent mappings, which take place in getting from the adjacency matrix to the equi-smooth graphon. The first of these linearly maps the adjacency matrix X -nodes $i \in \{1, \dots, m\}$ and Y -nodes $j \in \{1, \dots, l\}$, onto the anisotropic graphon bounded on $(0, 1)^2$. The second linearly stretches the anisotropic graphon by scale factor γ with respect to the X nodes, and $1/\gamma$ with respect to the Y -nodes, giving the equi-smooth graphon.

A co-blockmodel approximates the graphon, which requires that $k^{(X)}$ and $k^{(Y)}$ grow with m and l , respectively. However, the co-blockmodel also allows us to model the scenario in which $k^{(X)}$ and $k^{(Y)}$ grow at different rates with respect to m and l , i.e., that the number of X and Y -node groupings are unrelated. This might apply when co-clustering is used to analyse consumer-product review data, if e.g.,

an established supplier who is focussed on a certain type of products, expands into new markets. In that example, the increasing number of customers m , corresponds to a consumer demographic which has an increasing number of underlying groupings, $k^{(X)}$, but the increasing range of products, l , may reflect a much smaller increase in variety of the products offered, $k^{(Y)}$.

4 Identification and Comparison of Co-communities

Fitting the stochastic co-blockmodel by spectral clustering as described in Algorithm 1, involves using k -means to cluster the X and Y -nodes in the spaces of the left and right singular vectors of the co-Laplacian (equation 7). However, as will be subsequently illustrated, this leads to a problem of identifiability which does not arise when fitting the symmetric stochastic blockmodel to unipartite networks by spectral clustering. This problem of identifiability is precisely the question of estimating the set C (Definition 4) of pairings of X -node groupings $g^{(X)} \in G^{(X)}$ with Y -node groupings $g^{(Y)} \in G^{(Y)}$.

Fitting the symmetric blockmodel in the unipartite community detection setting, there are exactly $k = k^{(X)} = k^{(Y)}$ communities (because of symmetry). Each row grouping matches up with exactly one column grouping, because the row and column groupings are the same thing. On the other hand, fitting the asymmetric co-blockmodel by spectral clustering as in Algorithm 1 leads to $k^{(X)}$ and $k^{(Y)}$ row and column clusters. Hence, these $k^{(X)}$ and $k^{(Y)}$ row and column clusters provide $k^{(X)} \times k^{(Y)}$ potential co-communities. Which of these are significant? The best-known solution to this problem [Dhillon, 2001], instead of clustering the X and Y -nodes separately, instead normalises and concatenates the left and right singular vectors, and then clusters all the nodes at once. However, this approach has serious limitations: it again requires $k^{(X)} = k^{(Y)}$. Also, if the two types of nodes represent very different types of observations, then in practice we have found that method to perform less well. For this comparison, we define performance in terms of overlap with previously-defined groupings of nodes or variables.

So how should we assess and compare the $k^{(X)} \times k^{(Y)}$ potential co-communities, each of which is a different pairing of an estimated X -node grouping $\hat{g}^{(X)} \in \hat{G}^{(X)}$, with an estimated Y -node grouping $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$, to provide an assignment of the X -nodes and Y -nodes to co-communities, which is in some sense optimal? In practice, we expect the number of co-communities, $T = |C|$ (where $|\cdot|$ represents cardinality), to be significantly less than $k^{(X)} \times k^{(Y)}$. In the unipartite community detection setting, $k^{(X)} = k^{(Y)} = k$, and hence in effect there we have $T = k = \sqrt{k^{(X)} \times k^{(Y)}}$.

To estimate the set of co-communities, $c_t \in C$, $t = \{1, \dots, T\}$, in this bipartite network setting, we calculate the ‘local co-modularity’ for each pairing $\hat{g}^{(X)}$ with $\hat{g}^{(Y)}$, by considering a relevant sub-part of the co-modularity matrix \mathbf{B} (equation 15):

Definition 6 (Local co-modularity). *With \mathbf{A} given by Definition 1, with $\mathbf{d}^{(X)}$, $\mathbf{d}^{(Y)}$ and d^{++} given by Definition 4, with*

$$B_{ij} = A_{ij} - \frac{d_i^{(X)} d_j^{(Y)}}{d^{++}}, \quad \mathbf{B} = \mathbf{A} - \frac{1}{d^{++}} \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top, \quad (15)$$

and with the set of X -node groupings and the set of Y -node groupings estimated according to Algorithm 1 as $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$ respectively, where $|\hat{G}^{(X)}| = k^{(X)}$ and $|\hat{G}^{(Y)}| = k^{(Y)}$, where $|\cdot|$ represents cardinality, for a particular pairing of estimated X -node grouping $\hat{g}^{(X)} \in \hat{G}^{(X)}$ with estimated Y -node grouping $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$, the local co-modularity $Q_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)})$ is defined as:

$$Q_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)}) = \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij}. \quad (16)$$

Each of the $k^{(X)} \times k^{(Y)}$ possible pairings of $\hat{g}^{(X)}$ with $\hat{g}^{(Y)}$ can be defined, or not, as a co-community; doing so means that they are included in, or excluded from, the estimated set of co-communities \hat{C} (Definition 4). To consider all permutations, $2^{k^{(X)} \times k^{(Y)}}$ such assignments would need to be considered, which would be computationally very demanding. However, this problem can be avoided by defining summary statistics targeted for particular purposes. The three such purposes which we consider here are described in the following subsections: 4.1 Comparing potential co-communities and assessing their strength; 4.2 Arranging the co-communities for visualisation; 4.3 Defining an algorithmic objective function to be optimised, when determining co-community partitions.

4.1 Comparing and assessing significance of co-communities

Under a null model of no co-community structure, $\theta_{z^{(X)}(i), z^{(Y)}(j)} = \text{constant}$, for all i, j . Therefore, referring to the log-linear model [Perry and Wolfe, 2012], equation 1 becomes:

$$A_{ij} \sim \text{Bernoulli} \left(\frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}} \right), \quad (17)$$

where we have defined:

$$\theta_{z^{(X)}(i), z^{(Y)}(j)} = 1/\pi^{++}. \quad (18)$$

Hence under this null,

$$\mathbb{E}(A_{ij}) = \frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}} \implies \mathbb{E}(B_{ij}) = 0.$$

We define the informal idealised quantities $\tilde{\mathbf{B}}$ and \tilde{Q}_{XY} in comparison with equations 15 and 16:

$$\tilde{\mathbf{B}} = \mathbf{A} - \frac{1}{\pi^{++}} \boldsymbol{\pi}^{(X)} \left(\boldsymbol{\pi}^{(Y)} \right)^\top, \quad (19)$$

and

$$\tilde{Q}_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)}) = \frac{1}{\pi^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} \tilde{B}_{ij}, \quad (20)$$

where the empirical degree distributions $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ have been replaced by the theoretical node connectivity parameters $\boldsymbol{\pi}^{(X)}$ and $\boldsymbol{\pi}^{(Y)}$, and the empirical normalisation factor d^{++} is also replaced by the theoretical normalisation factor π^{++} .

If the pairing of X and Y -node groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ exhibit some co-community structure, then equation 18 no longer holds, and so the null model does not hold either. The stronger this co-community structure is, the further we move from the null model, and the greater θ becomes relative to $1/\pi^{++}$. This corresponds to $\mathbb{E}(A_{ij})$ becoming larger than $\pi_i^{(X)} \pi_j^{(Y)} / \pi^{++}$, which is equivalent to the observed number of edges in the co-community becoming greater than the expected, under the null of no co-community structure. This in turn means that \tilde{Q}_{XY} also becomes more positive. In other words, the further we move from the null model, the greater tendency of the X -nodes and Y -nodes of these groups to form connections with one another (compared with their expected propensity to make connections with any nodes, of the opposite type), and therefore constitute a strong co-community. Hence, a parsimonious method of comparing potential co-communities is simply to compare their local co-modularity, $Q_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)})$. This naturally leads to a ranking of potential co-communities according to their strength.

An estimate of statistical significance of a potential co-community can also be made, as follows. Noting that, with adjacency matrix \mathbf{A} defined according to the Bernoulli distribution of Definition 1, with fixed $\theta_{z^{(X)}(i), z^{(Y)}(j)} = 1/\pi^{++}$,

$$\text{Var}(\tilde{B}_{ij}) = \text{Var}(A_{ij}) = \left(\frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}} \right) \left(1 - \frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}} \right),$$

and assuming probabilities of observing links between different pairs of nodes are independent, the variance of $\tilde{Q}_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)})$ can be approximated as:

$$\text{Var}(\tilde{Q}_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)})) = \frac{1}{(\pi^{++})^2} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} \left(\frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}} \right) \left(1 - \frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}} \right), \quad (21)$$

where the factor $1/(\pi^{++})^2$ is due to the factor $1/(\pi^{++})$ in equation 20. Hence, assuming $\mathbf{d}^{(X)} \xrightarrow{p} \boldsymbol{\pi}^{(X)}$, $\mathbf{d}^{(Y)} \xrightarrow{p} \boldsymbol{\pi}^{(Y)}$ and $d^{++} \xrightarrow{p} \pi^{++}$, and assuming the potential co-community defined by $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ is comprised of sufficiently many nodes for a Gaussian approximation to hold, we can test the significance of $Q_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)})$ with a z -test, with zero mean and with $\text{Var}(Q_{XY})$ estimated as $\text{Var}(\tilde{Q}_{XY})$ in

equation 21, also replacing $\pi_i^{(X)}$ with $d_i^{(X)}$, $\pi_j^{(Y)}$ with $d_j^{(Y)}$ and π^{++} with d^{++} . A pairing $\hat{g}_p^{(X)}$ and $\hat{g}_q^{(Y)}$ is then defined as a co-community \hat{c} and included in \hat{C} (Definition 4), i.e., $\{p, q\} = \hat{c} \in \hat{C}$, if and only if this pairing $\hat{g}_p^{(X)}$ with $\hat{g}_q^{(Y)}$ is significant according to this z -test, at some significance level. We note that, in practice, this is only a rough approximation of significance, also because by specifying in advance the co-community node-groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$, we have introduced dependencies between the X and Y -nodes of this co-community.

4.2 Arranging the co-communities for visualisation

A standard task in exploratory data analysis using variants of the stochastic block model, is arranging the detected communities so they can be visualised in a helpful way. This visualisation is usually carried out by way of a heatmap representation of the adjacency matrix with the nodes grouped into communities. In the symmetric/unipartite community detection scenario, the communities occur along the leading diagonal of this ordered adjacency matrix. The communities themselves are often ordered along the leading diagonal according to their edge densities. In the bipartite co-community detection setting, co-communities may be present away from the leading diagonal, and there is no longer a restriction on how many co-communities a node may be part of - although we do not consider here the possibility of overlapping co-communities.

We propose then, that once the X -node groupings and Y -node groupings have been determined by spectral clustering as described above, a natural way to order these groups with respect to one another, is via row and column co-modularities, which we define as follows.

Definition 7. With d^{++} given by Definition 4, and with \mathbf{B} given by Definition 6, with the set of X -node groupings and the set of Y -node groupings estimated according to Algorithm 1 as $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$ respectively, the row and column modularities $Q_{\text{row}}(\hat{g}^{(X)})$ and $Q_{\text{column}}(\hat{g}^{(Y)})$ are defined, for $\hat{g}^{(X)} \in \hat{G}^{(X)}$ and $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$, as:

$$Q_{\text{row}}(\hat{g}^{(X)}) = \sum_{\hat{g}^{(Y)} \in \hat{G}^{(Y)}} \left| \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij} \right| \quad (22)$$

and

$$Q_{\text{column}}(\hat{g}^{(Y)}) = \sum_{\hat{g}^{(X)} \in \hat{G}^{(X)}} \left| \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij} \right|. \quad (23)$$

Considering the absolute values of the local co-modularities in these sums serves to prioritise the most extreme choices of divisions of nodes into co-communities, according to their local co-modularities. On the other hand if absolute values were not considered here, the row and column modularities would always be zero, because the rows and columns of \mathbf{B} must always sum to zero. The row and column co-modularities are the sums, respectively, of the absolute values of the local co-modularities along the rows and columns respectively, of the ordered adjacency matrix. Hence, they represent a measure of how extreme the co-community divisions are, in each row and column, according to the groupings defined by $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$. By ordering the X -node and Y -node groupings by decreasing $Q_{\text{row}}(\hat{g}^{(X)})$ and $Q_{\text{column}}(\hat{g}^{(Y)})$ respectively, co-communities with the largest local co-modularities will tend to congregate towards the top-left of the ordered adjacency matrix. This is a natural arrangement for visualisation as a heatmap, because it tends to place the strongest co-communities together in this corner, and so the attention is intuitively drawn to this region.

We note that there may be other equally effective ways of arranging the adjacency matrix for visualisation as a heatmap. However, this method is effective, and it is a parsimonious solution in the context of co-modularity, because row and column modularities are very simply and intuitively related to local co-modularity. In the case that there is no co-community structure present, such as under the null model of equation 17, then Q_{row} and Q_{column} as defined in Definition 7 would also tend to be close to zero, and the ordering would cease to be meaningful. However, if there are even a few significant co-communities present, their corresponding X and Y -node groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ would stand out, as assessed by Q_{row} and Q_{column} . Therefore these $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ would be placed at the top of the respective orderings, with the co-community pairings tending towards in the top-left corner. The other rows and columns,

which do not contain significant co-communities, would have corresponding Q_{row} and Q_{column} close to zero. Hence, these rows and columns would be naturally ordered according to their irrelevance. They would accordingly be placed further away from the top-left of the heatmap, giving the intuition that they are unimportant.

4.3 Defining an objective function for optimising the co-community partitions

Defining an objective function over the whole network, in terms of the assignments of the nodes to X -node and Y -node groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$, allows optimisation of these node assignments. It also provides a means of comparison of algorithmic parameters and other design choices in the practical implementation of the methods. It would be most ideal, for a trial assignment of nodes to $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$, to estimate the set of co-communities \hat{C} using the method of Section 4.1, and then to calculate the co-modularity according to Definition 4. However, for a large number of repetitions within an algorithm, or for an iterative search and optimisation, this would be computationally inefficient. Instead, we define the global co-modularity to be used as an objective function for such purposes, as follows:

Definition 8. With d^{++} given by Definition 4, and with \mathbf{B} given by Definition 6, with the set of X -node groupings and the set of Y -node groupings estimated according to Algorithm 1 as $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$ respectively, the global co-modularity is defined, for $\hat{g}^{(X)} \in \hat{G}^{(X)}$ and $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$, as:

$$Q_{global} = \sum_{\hat{g}^{(Y)} \in \hat{G}^{(Y)}} \sum_{\hat{g}^{(X)} \in \hat{G}^{(X)}} \left| \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij} \right|. \quad (24)$$

For a pairing $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$, the local co-modularity $Q_{XY}(\hat{g}^{(X)}, \hat{g}^{(Y)})$ represents the strength of the co-community structure in that grouping of X -nodes and Y -nodes. If the absolute value was not considered in the sum, Q_{global} would always be zero. Hence, by prioritising a sum of the absolute values of the local co-modularity of all pairings $\hat{g}^{(X)}$ with $\hat{g}^{(Y)}$, we prioritise an extreme division of the X -nodes and Y -nodes into co-communities, as measured by the local co-modularity. This therefore corresponds to an extreme partition in terms of co-community structure, as assessed by co-modularity.

Spectral clustering usually requires the nodes to be grouped in the spaces of the top singular vectors of the co-Laplacian, and this grouping is often carried out by k -means, as described in Algorithm 1. Because k -means optimisation is not convex, the converged result may be a local optimum. Hence, implementations of k -means often begin at a random start-point, with the optimisation run several times from random start-points, choosing the result which is in some sense optimal. In the community-detection setting, a natural statistic to maximise in this optimisation is the Newman-Girvan modularity. An equivalent statistic here to maximise in this co-community detection setting is hence the global co-modularity, which is intuitively linked to the local co-modularity measure of co-community structure. In the community-detection setting, assignments to communities can also be optimised by carrying out node-swapping between communities, in order to maximise the Newman-Girvan modularity [Blondel et al., 2008]. The global co-modularity is a statistic which could be equivalently maximised, in this co-community detection setting.

5 Examples

We now present results of applying the above methodology to simulated data, and to real data relating to movie reviews and to genomic patterns. We fit the degree-corrected stochastic co-blockmodel by spectral clustering, as detailed above, with the following additional practical details.

In the context of community detection, fitting the degree-corrected stochastic blockmodel using spectral clustering, when calculating the Laplacian it is advantageous to slightly inflate the degree distribution (regularisation) [Qin and Rohe, 2013], a trick which made Google’s original page-rank algorithm [Page et al., 1999] so effective in web-searching. Here in the co-community detection setting, correspondingly when calculating the co-Laplacian (equation 7), we inflate the diagonals of $\mathbf{D}^{(X)}$ and $\mathbf{D}^{(Y)}$ by the medians of $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$, respectively. Further, when fitting variants of the stochastic blockmodel by spectral clustering with k -means, nodes with small leverage score (which are usually low-degree nodes) can be

excluded from the k -means step [Qin and Rohe, 2013]; this practice is also followed here. We note that these regularisation steps have not previously been carried out in this co-community detection / co-clustering setting.

We also note that while spectral clustering is in general computationally intensive, binary adjacency matrices such as those dealt with in this setting tend to be very sparse. Further, we only require $k = \text{Max}(k^{(X)}, k^{(Y)})$ components of the singular value decomposition, a number which tends to be two or more orders of magnitude smaller than the maximum dimension of the adjacency matrix. Efficient computational methods exist to find the top few components in the singular value decomposition of large sparse matrices [Sørensen, 1992, Lehoucq and Sørensen, 1996], with implementations in *Matlab* and *R*, meaning that these methods are easy to implement and practical for large networks.

The k -means clustering algorithm begins with a random start-point, and hence it can provide a different result each time it is run. We therefore run the k -means step in the spectral clustering several times, choosing the result which maximises the global co-modularity (equation 24). We run k -means repeatedly until the output is visually assessed to have stabilised, at which point it can be seen from the convergence plot that there is very little, if any, improvement in co-modularity achieved by further runs of k -means. An example of such convergence in the genomics data set presented in Section 5.2 is shown in Figure 1.

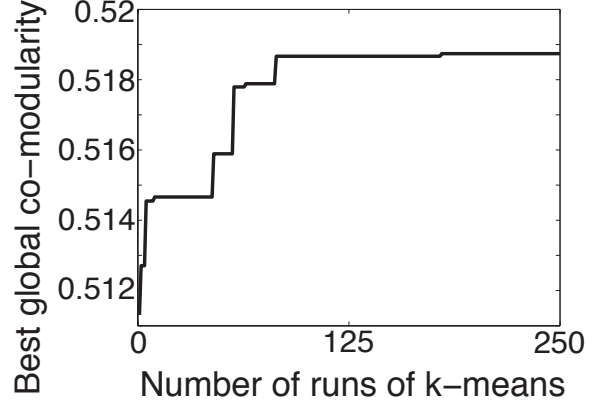


Figure 1: Convergence of the co-modularity.

The co-modularity converges well to a maximum, within 250 runs of k -means, in the genomics data set. For reference, the co-modularity is consistently found to be 0 when calculated based on randomly assigned co-community partitions of similar size.

5.1 Simulation Study

We carried out a simulation study, to evaluate the effectiveness of this co-community detection methodology against generated networks with known ground-truth co-communities. A classic generative model for exchangeable random networks with heterogeneous degrees is the logistic-linear model [Perry and Wolfe, 2012]. We use a version here for bipartite networks, with additional co-community structure, defined as:

$$\text{Logit}(p_{ij}) = \alpha_i^{(X)} + \alpha_j^{(Y)} + \theta_{ij},$$

where p_{ij} defines the probability of an edge being observed between nodes i and j . We choose to use this model, because the parameters can take any real values, and the edge probabilities p_{ij} will still be between 0 and 1. This model only deviates from the equivalent log model when the parameter values become very large, which is what prevents p_{ij} from reaching (and exceeding) 1 [Perry and Wolfe, 2012]. Further, the blockmodel approximates any smooth function, and hence the model can be used purely in the sense of approximation [Olhede and Wolfe, 2014, Choi et al., 2014]. The node-specific parameters $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ are elements of parameter vectors $\alpha^{(X)}$ and $\alpha^{(Y)}$ which define power law degree distributions for the X and Y -nodes. We would like power-law degree distributions for the nodes; this is a characteristic of scale-free networks [Barabási and Oltvai, 2004], which are found to be physically realistic in a wide range of scenarios, including biological networks [Wagner, 2002], and social networks [Barabási and Albert, 1999]. The parameters $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ are each generated as the logarithms of samples taken from a bounded Pareto distribution as in [Olhede and Wolfe, 2012]. We note that because $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ are chosen to be random, our generated networks are exchangeable [Kallenberg, 2005], whereas if $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ were defined deterministically, these networks would instead be generated under the inhomogeneous random graph model [Bollobás et al., 2007]. The co-community parameter θ_{ij} is allowed to take two values: $\theta_{ij} = \theta_{\text{in}}$ if i and j are in the same co-community, and $\theta_{ij} = \theta_{\text{out}}$ otherwise, which is equivalent to the modelling constraint we applied in equation 2. After generating the p_{ij} , the network is generated by sampling each A_{ij} ,

$$A_{ij} \sim \text{Bernoulli}(p_{ij}).$$

We note that this is the correctly specified model, for the co-community detection which we describe, and carry out in this study.

The co-communities themselves are planted in the network as randomly chosen groups of 150 of each type of node, with the maximum number of co-communities equal to $k^{(X)} \times k^{(Y)}$. By analogy with the unipartite/symmetric community detection setting, we choose to set the number of co-communities T as the square-root of this theoretical maximum, $T = \sqrt{k^{(X)} \times k^{(Y)}}$. As discussed in Section 4, in the unipartite community detection setting there is a constraint on the number of communities, $k = k^{(X)} = k^{(Y)}$, because the X -node and Y -node groupings are the same thing. This constraint does not exist in the bipartite co-community detection setting, and so the theoretical maximum number of co-communities is $k^{(X)} \times k^{(Y)}$, i.e., the square of the number of communities in the equivalent symmetric community detection setting. However, we expect the number of co-communities to be significantly less than this in practice, and so by default, we choose $T = \sqrt{k^{(X)} \times k^{(Y)}}$ as the number of co-communities, although we note that many other choices would also be valid here.

We test the methods on networks generated with $k^{(X)}$ and $k^{(Y)}$ ranging from 8 and 6 respectively up to 80 and 60 respectively (corresponding to values of numbers of nodes, m and l , ranging from 1200 and 800 up to 12000 and 8000, respectively). We also test the methods on networks generated with values of θ_{in} from 10 to 50, which corresponds to within co-community edge density $\rho_{\text{in}} \in \{0.039, 0.15, 0.34, 0.6\}$, and we set $\theta_{\text{out}} = 1$, corresponding to outside or between co-community edge density $\rho_{\text{in}} = 0.0013$. For each combination of parameters, we carry out 50 repetitions of network generation and co-community detection, to enable assessment of the variability of the accuracy of the co-community detection (with more repetitions, the computational cost becomes prohibitive).

After generating the networks, we detect co-communities according to the methods described above, based on the same values of $k^{(X)}$ and $k^{(Y)}$ that we used to generate the networks. We keep these values the same, to understand specifically how the co-community detection methodology is working. This means there are $k^{(X)} \times k^{(Y)}$ potential co-communities, and we assess each in terms of strength and significance, as discussed in Section 4.1. Hence, we define the estimated set of co-communities \hat{C} , as all combinations of detected X and Y -node groupings $\hat{g}^{(X)} \in \hat{G}^{(X)}$ with $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$ which are significant according to a z -test with zero mean and variance calculated as in Equation 21. We define significance according to FDR (false discovery rate) corrected [Benjamini and Hochberg, 1995] p -value < 0.05 . This tends to result in more co-communities being detected than were originally planted (primarily due to some being split), however we note that the main aim of this methodology is to find a good representation of the underlying co-community structure (as assessed by co-modularity), rather than to reproduce it exactly.

To compare detected co-communities with the ground-truth planted co-communities, we use the normalised mutual information (NMI) [Danon et al., 2005]. The NMI compares the numbers of nodes which appear together in the found co-communities, compared with whether they appeared together in the planted co-communities (adjusted for group sizes). It has been used previously in the co-community detection context [Larremore et al., 2014], as well as the unipartite community-detection context [Zhao et al., 2012]. The NMI takes the value 1 if the co-communities are perfectly reproduced in the co-community detection, and 0 if they are not reproduced at all, and somewhere in between if they are partially reproduced. The results, together with examples of randomly generated adjacency matrices, are shown in Figure 2, which shows that the method performs well as long as there is sufficient within-co-community edge density, and performs well as the number of co-communities increases.

5.2 Genomics Data-set

We present an example of a practical application of these methods to a challenging problem in genomics. A gene encodes how to make a gene product, and the corresponding gene expression level quantifies how much gene product is currently being produced. Hence, the gene expression level indicates the extent to which a gene is ‘active’, or ‘switched on’. DNA methylation is a gene regulatory pattern, meaning that it influences the activity/expression level of particular genes. DNA methylation patterns are themselves influenced by the expression levels/activity of other genes. However, much is still unknown about the interaction between DNA methylation patterns and gene expression patterns [Jones, 2012]. It is of much interest to uncover groups of genes with methylation patterns which are linked to the expression patterns of other groups of genes, to allow biological hypotheses to be formed, which can then be investigated further, experimentally and computationally. Hence, this is a natural scenario to be approached with co-

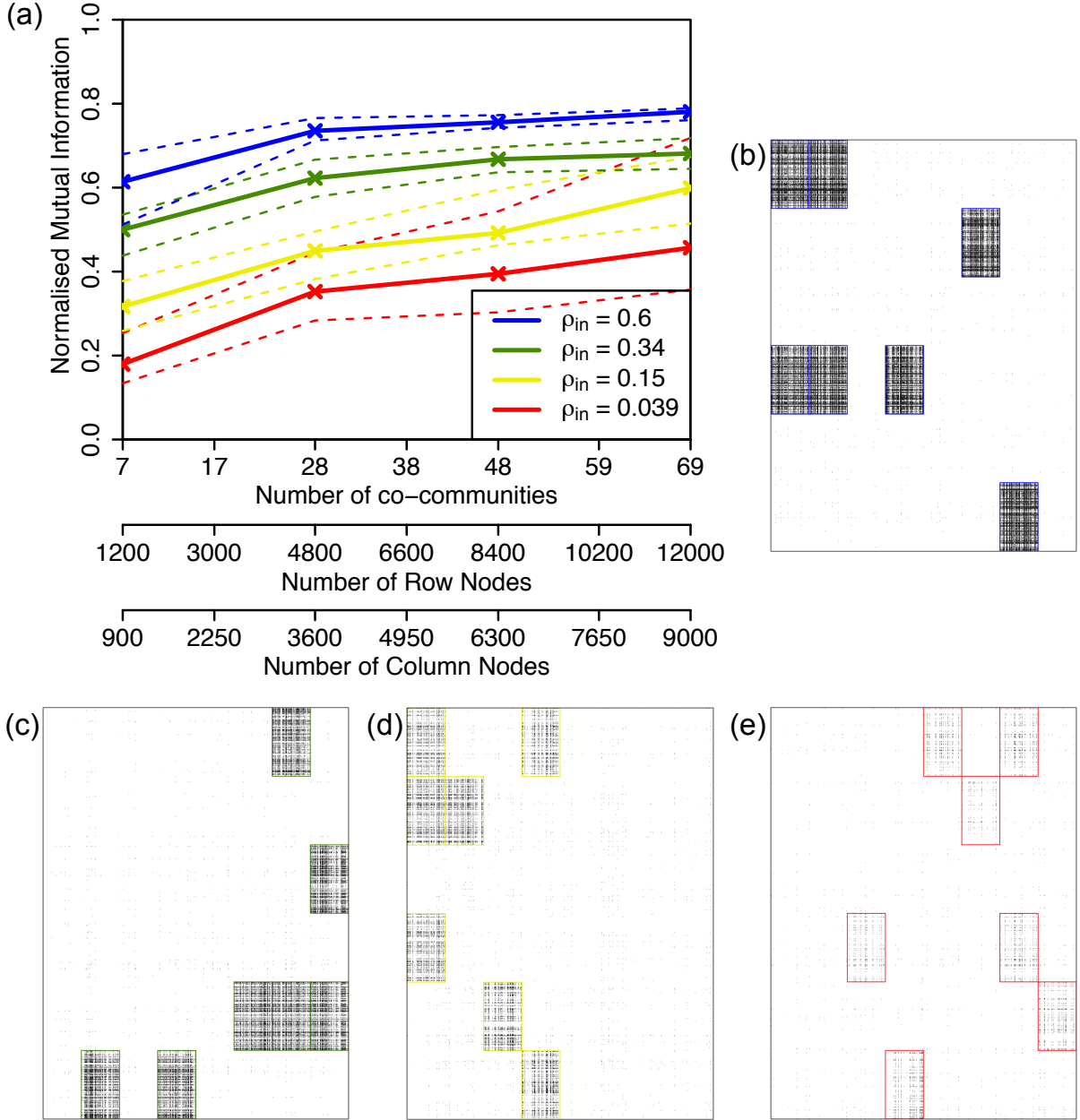


Figure 2: Simulation Study.

(a) Normalised mutual information (NMI) compares detected co-communities with ground-truth planted co-communities. (b)-(e) Examples of generated networks all with $nR = 1200$, $nC = 900$, $kR = 8$, $kC = 6$, and 7 planted co-communities; entries in the adjacency matrix equal to 1 (representing a network edge) are marked in black; planted co-communities are outlined in colour. (b) $\theta_{in} = 40$, within-community edge density $\rho_{in} = 0.6$; (c) $\theta_{in} = 30$, $\rho_{in} = 0.34$; (d) $\theta_{in} = 20$, $\rho_{in} = 0.15$; (e) $\theta_{in} = 10$, $\rho_{in} = 0.039$. For all networks, $\theta_{out} = 1$, outside/between co-community edge density $\rho_{out} = 0.0013$

community detection, as the method offers the potential to uncover latent structure not easily identifiable otherwise.

As a measure of the DNA methylation (DNAm) pattern of each gene, we choose to consider here intra-gene DNA methylation variability (IGV), as it is a per-gene measure of DNA methylation variance which has been shown to be strongly associated with disease [Bartlett et al., 2013]. We denote the gene expression variables $X(i)$, $i = 1, \dots, m$ and the DNAm variables $Y(j)$, $j = 1, \dots, l$; i.e., $X(i)$ and $Y(j)$ refer to the measurements for particular genes of gene expression and DNA methylation IGV respectively. We define a network edge $A_{ij} = 1$ if variables $X(i)$ and $Y(i)$ are significantly correlated, and we set $A_{ij} = 0$

otherwise.

We carried out co-community detection on this genomics data set according to the methods described above (data source: The Cancer Genome Atlas [Hampton, 2006], breast cancer invasive carcinoma data set, basal tumour samples only). Figure 3(a) shows the adjacency matrix after carrying out co-community detection, ordering the X and Y -node groupings by row and column co-modularity (equations 22 and 23). Figure 3(b) (inlay) shows the same adjacency matrix ordered along its margins alphabetically by gene name, i.e., without ordering the margins using co-community detection. Hence, Figure 3(b) shows a baseline in which the nodes are essentially randomly ordered, against which to compare the adjacency matrix after co-community detection, and ordering based upon it. The co-community structure is clearly revealed in Figure 3(a), whereas no co-community structure is visible in Figure 3(b). We define a co-community $\hat{c} \in \hat{C}$ as a combination of X -node grouping $\hat{g}^{(X)} \in \hat{G}^{(X)}$ with Y -node grouping $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$ which is significant according to a z -test with zero mean and variance calculated as in equation 21, with significance defined by FDR-corrected p -value < 0.05 . The numbers of X and Y -node groupings, $k^{(X)}$ and $k^{(Y)}$, are estimated according to equations 12 and 13 as 89 and 67 respectively, leading to 5963 potential co-communities, of which $\hat{T} = 2018$ are found to be significant. We tested these 2018 significant co-communities for domain relevance, by testing the overlap of the genes (nodes) of each co-community, separately with each of 10295 known gene-groups (data source: <http://www.broadinstitute.org/gsea/msigdb/>). This type of analysis is often called ‘gene set enrichment analysis’ (GSEA) [Subramanian et al., 2005]. We found that 1340 (66%) overlap significantly (Fisher’s exact test, FDR-adjusted $p < 0.05$) with these known gene-sets, confirming the domain relevance of this result, as well as indicating novel findings which could be investigated further by experimental biologists.

5.3 Movie-review Data-set

We present a second, contrasting example of a practical application of these methods to real data, to a consumer-product review dataset. We downloaded movie review data from the *Movie Lens* database, which details 1 000 209 reviews of 3952 different movies, by 6040 unique users who each provided at least 20 different reviews (data source: <http://grouplens.org/datasets/movielens/>). Denoting movies by the variables $X(i)$, $i = 1, \dots, m$ and users by the variables $Y(j)$, $j = 1, \dots, l$, we define a network edge, i.e., $A_{ij} = 1$, if movie $X(i)$ has been reviewed by user $Y(j)$, and no edge, i.e., $A_{ij} = 0$, otherwise. Covariate information is also available, assigning each movie to one of 18 categories, and classifying each user into one of 7 age groups and 20 professions.

The numbers of X and Y -node groupings, $k^{(X)}$ and $k^{(Y)}$, are estimated as 77 and 95 respectively according to equations 12 and 13. However, the granularity of these estimates is much greater than that of the covariate information we have available for verification of detected co-clusters. Because these estimates of $k^{(X)}$ and $k^{(Y)}$ represent the maximum granularity for the stochastic blockmodel to remain consistent, we are free to reduce $k^{(X)}$ and $k^{(Y)}$ (although we note that this will not correspond to the optimal choice in terms of minimising the MISE). We reduce $k^{(X)}$ and $k^{(Y)}$ to a similar granularity to that of the covariate information which we have available. Hence, for co-community detection we take $k^{(X)} = 10$ and $k^{(Y)} = 15$, which also takes account of the fact that there is overlap between many of the covariate classes. Co-community detection was carried out using the methods described above, and Figure 4 shows the result. Of the 150 potential co-communities, $\hat{T} = 41$ are found to be significant using the z -test as before, and are thus defined as co-communities. The X and Y -nodes of these 41 co-communities are tested for overlap with the covariate groups. Of these, 22 are found to overlap significantly with one or more covariate groups (Fisher’s exact test, FDR-corrected $p < 0.05$), and these are highlighted in red, with the significant covariate groups shown along the margins, in Figure 4. Many of the findings are predictable: horror, sci-fi and war films tend to be watched by younger people; drama and romance are popular across the board. Others need more explanation, for example a group of children’s movies and musicals tend to be reviewed by 25-34 year old customer service professionals. However, we can expect that this is a demographic group of people who tend to have younger children who they watch movies with. Other children’s movies are grouped together with animation, fantasy and horror, and tend to be watched by both younger and older groups. This might reflect very broad classifications used for such movies, many of which in reality could be fairly similar. Also that these are groups of people who would tend to watch movies together. An important conclusion to draw, is that the covariate information available for this data-set appears to be unspecific in comparison with the detail which can be revealed by these co-community detection methods.

(a)

DNAm IGV

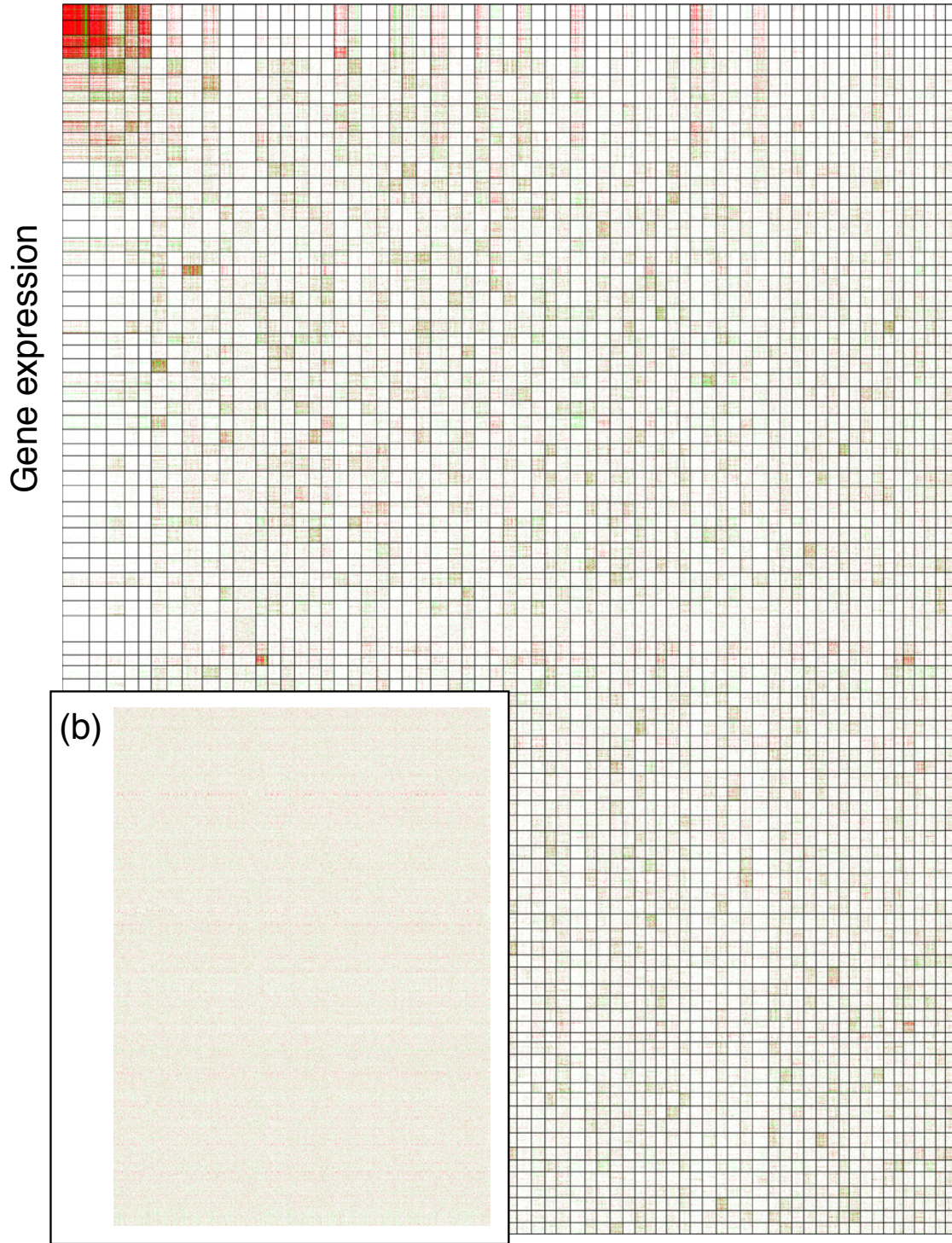


Figure 3: Co-communities in the genomics data set.

(a) Genes are ordered along the margins of the adjacency matrix, according to co-communities detected by the methods presented here. Partitions between detected co-communities are shown with black lines. (b) The same adjacency matrix ordered along its margins alphabetically by gene name, i.e., without ordering the margins using co-community detection. Entries in the adjacency matrix equal to 1 (representing a network edge) are coloured, with green and red indicating positive and negative associations, respectively. N.B., this colour-scheme contrasts with that used in Figure 4, in which network edges do not have associated signs, and hence are all coloured blue.

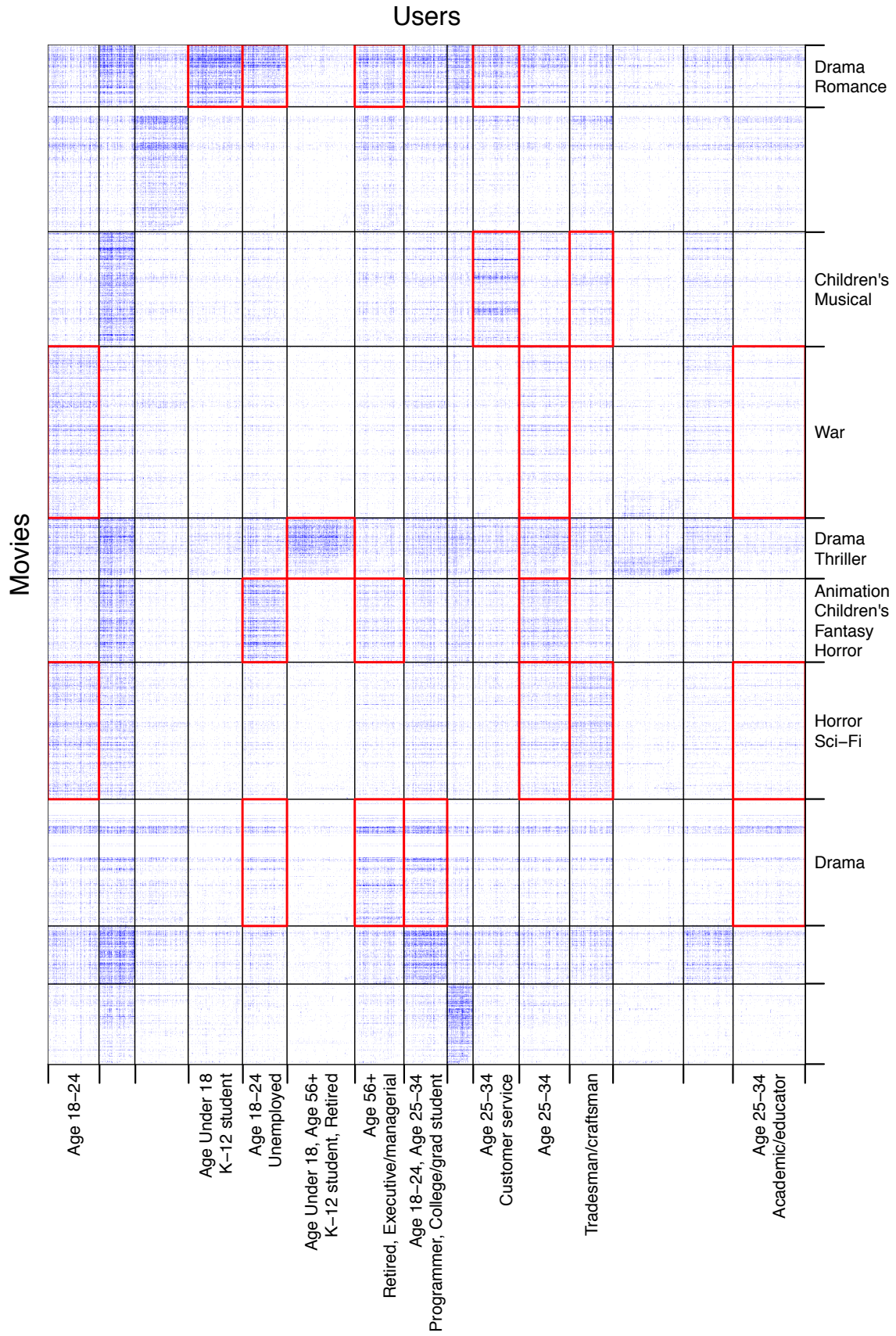


Figure 4: Co-communities in the movie-review data set.
 Entries in the adjacency matrix equal to 1 (representing a network edge) are coloured blue, and detected communities are outlined in black.

6 Conclusion

We have introduced the notion of co-modularity. We have shown how it can be used to perform co-community detection in bipartite networks, and how it fits with the notion of the stochastic co-blockmodel. We have shown how co-modularity can be used to compare co-communities, to calculate their strength and significance, to arrange them for visualisation, and to calculate an algorithmic objective function for optimisation. We have introduced the anisotropic graphon class, and have shown how to use it to estimate the optimum number of groups into which to divide the two types of nodes. We have also shown how this estimation can be simplified in certain circumstances for a more parsimonious scheme, and how to test whether this simplification is justified. We have addressed practical points about the implementation of the methodology, and have demonstrated its utility with a simulation study and application to two contrasting examples of real datasets, from genomics and consumer-product reviews.

An interesting extension to this methodology would be to consider overlapping blocks in the stochastic co-blockmodel, a problem which has already been successfully addressed in the context of the stochastic blockmodel for unipartite networks [Latouche et al., 2011], and in co-clustering without fitting the stochastic blockmodel [Madeira and Oliveira, 2004]. Another interesting application would be to develop an online version of the method as a computationally efficient approach to large and growing data-sets [Zanghi et al., 2010].

This methodology would be expected to work similarly well in many other contexts, such as interpersonal networks where the individuals are of two distinct categories, such as teachers and students, or publication networks where the two types of variables are authors and papers. This methodology could also be expected to work in even more general settings of bi-clustering or co-clustering, in which the variables being clustered together are simply correlated, rather than having any tangible interactive behaviour in the real world. These methods are based on commonly available computationally efficient methods for large sparse matrices, and perform well on large datasets, with large numbers of co-communities, often performing better than methods based on model likelihoods.

References

- E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.
- D. J. Aldous. *Exchangeability and related topics*. Springer, 1985.
- A. A. Amini and E. Levina. On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*, 2014.
- A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.
- A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
- T. E. Bartlett, A. Zaikin, S. C. Olhede, J. West, A. E. Teschendorff, and M. Widschwendter. Corruption of the intra-gene DNA methylation architecture is a hallmark of cancer. *PloS One*, 8(7):e68285, 2013.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

- B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.
- T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *arXiv preprint arXiv:1404.6000*, 2014.
- S. H. Chan and E. M. Airoldi. A consistent histogram estimator for exchangeable graph models. *arXiv preprint arXiv:1402.1888*, 2014.
- D. Choi, P. J. Wolfe, et al. Co-clustering separately exchangeable network data. *The Annals of Statistics*, 42(1):29–63, 2014.
- L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.
- I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.
- P. Diaconis. Finite forms of de finetti’s theorem on exchangeability. *Synthese*, 36(2):271–281, 1977.
- T. Duong and M. Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15(1):17–30, 2003.
- et al.. E. M. Airoldi. Mixed membership stochastic blockmodels. *J. MLR*, 9:1981–2014, 2008.
- C. J. Flynn and P. O. Perry. Consistent biclustering. *arXiv preprint arXiv:1206.6927*, 2012.
- M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- P. K. Gopalan and D. M. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110:14534–14539, 2013.
- T. Hampton. Cancer genome atlas. *JAMA: The Journal of the American Medical Association*, 296(16):1958–1958, 2006.
- P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
- R. A. Horn and C. R. Johnson. Topics in matrix analysis. *Cambridge UP, New York*, 1991.
- P. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012.
- O. Kallenberg. *Probabilistic symmetries and invariance principles*, volume 9. Springer, 2005.
- D. B. Larremore, A. Clauset, and A. Z. Jacobs. Efficiently inferring community structure in bipartite networks. *arXiv preprint arXiv:1403.2933*, 2014.
- P. Latouche, E. Birmelé, C. Ambroise, et al. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.
- R. B. Lehoucq and D. C. Sørensen. Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821, 1996.
- S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45, 2004.
- M. Newman. Spectral methods for network community detection and graph partitioning. *arXiv preprint arXiv:1307.7729*, 2013.
- M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

- S. C. Olhede and P. J. Wolfe. Degree-based network models. *arXiv preprint arXiv:1211.6537*, 2012.
- S. C. Olhede and P. J. Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014. doi: 10.1073/pnas.1400374111. URL <http://www.pnas.org/content/111/41/14722.abstract>.
- L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.
- P. O. Perry and P. J. Wolfe. Null models for network data. *arXiv preprint arXiv:1201.5871*, 2012.
- T. Qin and K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.
- M. A. Riolo and M. Newman. First-principles multiway spectral partitioning of graphs. *arXiv preprint arXiv:1209.5969*, 2012.
- K. Rohe and B. Yu. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*, 2012.
- K. Rohe, S. Chatterjee, B. Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.
- D. C. Sørensen. Implicit application of polynomial filters in ak-step arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 13(1):357–385, 1992.
- A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545, 2005.
- A. Wagner. Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Research*, 12(2):309–315, 2002.
- M. Wand and M. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528, 1993.
- J. D. Wilson, S. Wang, P. J. Mucha, S. Bhamidi, and A. B. Nobel. A testing based extraction algorithm for identifying significant communities in networks. *arXiv preprint arXiv:1308.0777*, 2013.
- P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.
- H. Zanghi, F. Picard, V. Miele, C. Ambroise, et al. Strategies for online inference of model-based clustering in large and growing networks. *The Annals of Applied Statistics*, 4(2):687–714, 2010.
- Y. Zhao, E. Levina, J. Zhu, et al. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

7 Supplement

Supplement A: Derivation relating to Algorithm 1

Define $m, l, \mathbf{A}, \mathbf{B}, \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}, d^{++}, g^{(X)}, g^{(Y)}, k^{(X)}, k^{(Y)}, \Psi$ and Q_{XY} according to Definitions 1 - 6. Specify that $k^{(X)} = k^{(Y)} = 2$, that $T = 2$, that $c_1 = \{1, 1\}$, and that $c_2 = \{2, 2\}$; i.e., that there are two co-communities, the first of which consists of $g_1^{(X)}$ paired with $g_1^{(Y)}$, and the second of which consists of $g_2^{(X)}$ paired with $g_2^{(Y)}$. Define co-community label vectors \mathbf{s} and \mathbf{r} for the X and Y -nodes respectively, such that:

$$s_i = \begin{cases} 1, & \text{if } X\text{-node } i \text{ is in co-community 1,} \\ -1, & \text{if } X\text{-node } i \text{ is in co-community 2,} \end{cases} \quad (25)$$

and

$$r_j = \begin{cases} 1, & \text{if } Y\text{-node } j \text{ is in co-community 1,} \\ -1, & \text{if } Y\text{-node } j \text{ is in co-community 2.} \end{cases} \quad (26)$$

Hence:

$$\Psi(C; G^{(X)}, G^{(Y)}; i, j) = \frac{1}{2} (s_i r_j + 1),$$

and

$$Q_{XY} = \frac{1}{2d^{++}} \sum_{i=1}^m \sum_{j=1}^l B_{ij} (s_i r_j + 1).$$

Note that the rows of \mathbf{B} sum to zero:

$$\sum_{j=1}^l B_{ij} = \sum_{j=1}^l A_{ij} - \frac{d_i^{(X)}}{d^{++}} \sum_{j=1}^l d_j^{(Y)} = d_i^{(X)} - \frac{d_i^{(X)}}{d^{++}} \cdot d^{++} = 0.$$

Also, the columns of \mathbf{B} also sum to zero, by a similar argument. Hence:

$$Q_{XY} = \frac{1}{2d^{++}} \sum_{i=1}^m \sum_{j=1}^l B_{ij} s_i r_j. \quad (27)$$

When [Newman, 2013] derives the properties of unipartite network community detection he relaxes the constraint that the co-community labels take the values of ± 1 , to be able to arrive at an algorithmic solution. Nodes are then assigned to one community or the other, according to their sign (in the two-community scenario). A similar relaxation is made here, allowing $s_i \in \mathbb{R}$ and $r_j \in \mathbb{R}$, subject also to the following elliptical constraints, which allow for degree heterogeneity as in the degree corrected stochastic blockmodel:

$$\sum_{i=1}^m d_i^{(X)} s_i^2 = d^{++}, \quad (28)$$

$$\sum_{j=1}^l d_j^{(Y)} r_j^2 = d^{++}. \quad (29)$$

In the extreme scenario, in which $s_i \in \{-1, 1\}$ and $r_j \in \{-1, 1\}$, these constraints are equivalent to $d^{++} = \sum_{i=1}^m d_i^{(X)} = \sum_{j=1}^l d_j^{(Y)}$ (i.e., as per definition 4). This relaxation is equivalent to saying that nodes may be partly in one group, and partly in another group. N.B., ultimately each node will be assigned entirely to only the group it is most strongly associated with (according to s_i or r_j), and hence mixed membership does not occur in the final assignment of nodes to groups. For homogenous degree distributions, the constraints of equation 28 and 29 prevent the co-modularity from becoming arbitrarily large, as nodes are assigned many times over to many groups. For heterogenous degree distributions, the effect of the constraint is equivalent, except that the constraint is weighted to give importance to

high-degree nodes. This is achieved by the constraints of equation 28 and 29 restricting the weighted sum of the degrees (weighted by the assignment of nodes to groups) to be the equal to the total number of edges.

We wish to find the community assignment vectors \mathbf{r} and \mathbf{s} which maximise the co-modularity, i.e., we want to maximise Q_{XY} with respect to both \mathbf{r} and \mathbf{s} . To do this, we employ the Lagrange multipliers λ and μ , and equate the derivatives to zero, N.B., the partial derivatives with respect to $s_{i'}$ and $r_{j'}$ are used as the derivatives are taken with respect to these individual $i' \in \{1, \dots, l\}$, and $j' \in \{1, \dots, m\}$.

$$\begin{aligned} \frac{\partial}{\partial s_{i'}} \left[\sum_{i=1}^m \sum_{j=1}^l B_{ij} s_i r_j - \lambda \sum_{i=1}^m d_i^{(X)} s_i^2 - \mu \sum_{j=1}^l d_j^{(Y)} r_j^2 \right] &= 0, \\ \text{and } \frac{\partial}{\partial r_{j'}} \left[\sum_{i=1}^m \sum_{j=1}^l B_{ij} s_i r_j - \lambda \sum_{i=1}^m d_i^{(X)} s_i^2 - \mu \sum_{j=1}^l d_j^{(Y)} r_j^2 \right] &= 0, \\ \implies \sum_{j=1}^l B_{ij} r_j - 2\lambda d_i^{(X)} s_i &= 0, \end{aligned} \quad (30)$$

$$\text{and } \sum_{i=1}^m B_{ij} s_i - 2\mu d_j^{(Y)} r_j = 0. \quad (31)$$

Hence, taking $\mathbf{D}^{(X)}$ and $\mathbf{D}^{(Y)}$ as the diagonal matrices with the degree vectors $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ respectively on their leading diagonals,

$$\mathbf{B}\mathbf{r} = 2\lambda\mathbf{D}^{(x)}\mathbf{s} \quad (32)$$

$$\text{and } \mathbf{B}^\top\mathbf{s} = 2\mu\mathbf{D}^{(y)}\mathbf{r}. \quad (33)$$

Substituting for \mathbf{s} , equation 33 in 32, gives:

$$\left(\mathbf{D}^{(y)}\right)^{-1}\mathbf{B}^\top\left(\mathbf{D}^{(x)}\right)^{-1}\mathbf{B}\mathbf{r} = 4\lambda\mu\mathbf{r}, \quad (34)$$

$$\begin{aligned} \implies \left(\mathbf{D}^{(y)}\right)^{-1/2}\mathbf{B}^\top\left(\mathbf{D}^{(x)}\right)^{-1/2}\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}\mathbf{r} &= 4\lambda\mu\mathbf{r}, \\ \implies \left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)^\top\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)\mathbf{r} &= 4\lambda\mu\mathbf{r}, \end{aligned} \quad (35)$$

$$\implies \mathbf{M}^\top\mathbf{M}\mathbf{r} = 4\lambda\mu\mathbf{r}, \quad (36)$$

where

$$\mathbf{M} = \left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}.$$

By an identical argument, substituting 32 in 33 and re-arranging equivalently,

$$\mathbf{M}\mathbf{M}^\top\mathbf{s} = 4\lambda\mu\mathbf{s}. \quad (37)$$

Hence, \mathbf{s} and \mathbf{r} are eigenvectors of $\mathbf{M}\mathbf{M}^\top$ and $\mathbf{M}^\top\mathbf{M}$ respectively, with $4\lambda\mu$ the corresponding eigenvalue in both cases. Therefore, \mathbf{s} and \mathbf{r} are left and right singular vectors respectively of:

$$\mathbf{M} = \left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2},$$

with corresponding singular value $2\sqrt{\lambda\mu}$.

Multiplying equation 30 by $s_i/2d^{++}$, summing over i and referring to equation 28 gives:

$$\frac{1}{2d^{++}} \sum_{i=1}^m \sum_{j=1}^l B_{ij} s_i r_j = \frac{2\lambda}{2d^{++}} \sum_{i=1}^m d_i^{(X)} s_i^2 = \frac{2\lambda \cdot d^{++}}{2d^{++}} = \lambda,$$

hence referring to equation 27, we get:

$$Q_{XY} = \lambda. \quad (38)$$

Then equivalently multiplying equation 31 by $r_j/2d^{++}$, summing over j and referring to equation 29, and then referring to equation 27 gives:

$$Q_{XY} = \mu. \quad (39)$$

Therefore, referring again to equations 36 and 37, the maximum modularity solution is for the left and right singular vectors of M which correspond to the greatest singular value 2λ .

Now substituting equation 15 in equation 33, we get:

$$\begin{aligned} \mathbf{s}^\top \left(\mathbf{A} - \frac{1}{d^{++}} \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top \right) &= 2\mu \mathbf{r}^\top \mathbf{D}^{(y)}, \\ \implies \mathbf{s}^\top \mathbf{A} &= \frac{1}{d^{++}} \mathbf{s}^\top \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top + 2\mu \mathbf{r}^\top \mathbf{D}^{(y)}. \end{aligned} \quad (40)$$

Post-multiplying equation 40 by $\mathbf{1} = (1, 1, 1, \dots)$ leads to:

$$\begin{aligned} \mathbf{s}^\top \mathbf{d}^{(X)} &= \frac{1}{d^{++}} \mathbf{s}^\top \mathbf{d}^{(X)} \cdot d^{++} + 2\mu \mathbf{r}^\top \mathbf{d}^{(Y)} \\ \therefore \mu \mathbf{r}^\top \mathbf{d}^{(Y)} &= 0. \end{aligned}$$

Assuming that there is co-community structure present in \mathbf{A} , there must be positive co-modularity, i.e., $Q_{XY} > 0 \implies \mu > 0$ (referring back to equation 39), and therefore $\mathbf{r}^\top \mathbf{d}^{(Y)} = 0$. By an identical argument, also $\mathbf{s}^\top \mathbf{d}^{(X)} = 0$. Therefore, for eigenvectors \mathbf{r} corresponding to $Q_{XY} > 0$,

$$\mathbf{B}\mathbf{r} = \left(\mathbf{A} - \frac{1}{d^{++}} \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top \right) \mathbf{r} = \mathbf{A}\mathbf{r}$$

and so to find these eigenvectors with Q_{XY} maximised, instead of equation 35 we can consider

$$\begin{aligned} \left(\left(\mathbf{D}^{(x)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(y)} \right)^{-1/2} \right)^\top \left(\left(\mathbf{D}^{(x)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(y)} \right)^{-1/2} \right) \mathbf{r} \\ = (2\lambda)^2 \mathbf{r} \end{aligned} \quad (41)$$

which, referring back to equation 7, can be written in terms of the co-Laplacian \mathbf{L}_{XY} as:

$$\mathbf{L}_{XY}^\top \mathbf{L}_{XY} \mathbf{r} = (2\lambda)^2 \mathbf{r}.$$

By identical argument, we can also write:

$$\begin{aligned} \left(\left(\mathbf{D}^{(x)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(y)} \right)^{-1/2} \right) \left(\left(\mathbf{D}^{(x)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(y)} \right)^{-1/2} \right)^\top \mathbf{s} \\ = (2\lambda)^2 \mathbf{s} \end{aligned} \quad (42)$$

and

$$\mathbf{L}_{XY} \mathbf{L}_{XY}^\top \mathbf{s} = (2\lambda)^2 \mathbf{s}.$$

Hence, the co-Laplacian \mathbf{L}_{XY} has left and right singular vectors \mathbf{s} and \mathbf{r} respectively, with corresponding singular values 2λ . It can be seen that equation 41 has the eigenvector $\mathbf{1} = (1, 1, 1, \dots)$, as follows:

$$\begin{aligned} \left(\left(\mathbf{D}^{(x)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(y)} \right)^{-1/2} \right)^\top \left(\left(\mathbf{D}^{(x)} \right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(y)} \right)^{-1/2} \right) \mathbf{1} &= (2\lambda)^2 \mathbf{1} \\ \implies \left(\mathbf{D}^{(y)} \right)^{-1} \mathbf{A}^\top \left(\mathbf{D}^{(x)} \right)^{-1} \mathbf{A} \mathbf{1} &= (2\lambda)^2 \mathbf{1} \\ \implies \left(\mathbf{D}^{(y)} \right)^{-1} \mathbf{A}^\top \left(\mathbf{D}^{(x)} \right)^{-1} \mathbf{d}^{(X)} &= (2\lambda)^2 \mathbf{1} \\ \implies \left(\mathbf{D}^{(y)} \right)^{-1} \mathbf{A}^\top \mathbf{1} &= (2\lambda)^2 \mathbf{1} \\ \implies \left(\mathbf{D}^{(y)} \right)^{-1} \mathbf{d}^{(Y)} &= (2\lambda)^2 \mathbf{1} \\ \mathbf{1} &= (2\lambda)^2 \mathbf{1} \end{aligned}$$

and hence the corresponding eigenvalue is $(2\lambda)^2 = 1$, which by the Perron-Frobenius theorem, must be the greatest eigenvalue [Hom and Johnson, 1991, Newman, 2013]. An identical argument can also be applied to \mathbf{s} in equation 42. This means that the greatest singular value $2\lambda = 1$ corresponds to these left and right singular vectors which are both $\mathbf{1}$ (of lengths m and l respectively), however such singular vectors do not satisfy $\mathbf{r}^\top \mathbf{d}^{(Y)} = 0$ and $\mathbf{s}^\top \mathbf{d}^{(X)} = 0$. Therefore, to maximise the co-modularity in the case of two co-communities, we should divide the X and Y -nodes according to the left and right singular vectors respectively which correspond to the second greatest singular value.

The above explains how Algorithm 1 works for the case of two co-communities. An equivalent extension to k communities has been made In the unipartite community detection setting [Riolo and Newman, 2012]. To do so, the community labels are identified with the vertices of $k - 1$ simplices, i.e., for detection of 3 communities, the co-community labels would be the vertices of a triangle. Relaxing constraints equivalent to equations 28 and 29 means allowing the nodes to move away from the vertices of the simplex. This amounts to clustering the nodes in the space of the eigenvectors corresponding to the 2nd to k^{th} greatest eigenvalues of the Laplacian \mathbf{L} . This clustering is conventionally done using k -means. The reader is referred to [Riolo and Newman, 2012] for the detailed technical derivations relating to this. A similar extension can naturally be made in this co-community detection setting. To detect $k^{(X)}$ X -node groupings, and $k^{(Y)}$ Y -node groupings, the X and Y -nodes can be separately clustered (using k -means independently for the X and Y -nodes) in the spaces of the left and right singular vectors (respectively) corresponding to the 2nd to $k^{(X)\text{th}}$ and 2nd to $k^{(Y)\text{th}}$ greatest singular values, respectively, of the singular value decomposition of the co-Laplacian \mathbf{L}_{XY} .

Supplement B: Proof of Proposition 1

For the case of two co-communities, with θ_{in} and θ_{out} defined according to equation 2, with the co-community labels r_i and s_j defined as in Supplement A / section 7 (equations 25 and 26), and with $G^{(X)}$ and $G^{(Y)}$ defined according to Definition 1, we note (equivalently to [Newman, 2013]) that:

$$\theta_{z^{(X)}(i), z^{(Y)}(j)} = \frac{1}{2} (\theta_{\text{in}} + \theta_{\text{out}} + r_i s_j (\theta_{\text{in}} - \theta_{\text{out}})), \quad (43)$$

$$\text{and } \ln(\theta_{z^{(X)}(i), z^{(Y)}(j)}) = \frac{1}{2} \left(\ln(\theta_{\text{in}} \theta_{\text{out}}) + r_i s_j \ln\left(\frac{\theta_{\text{in}}}{\theta_{\text{out}}}\right) \right), \quad (44)$$

N.B., equations 43 and 44 only hold because $s_i \in \{-1, 1\}$ and $r_j \in \{-1, 1\}$. Substituting equations 43 and 44 into equation 3, and estimating the node-specific connectivity parameters $\boldsymbol{\pi}^{(X)}$ and $\boldsymbol{\pi}^{(Y)}$ by the degree distributions $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ leads to the profile likelihood:

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}) &= \sum_{i=1}^m \sum_{j=1}^l \left[\frac{A_{ij}}{2} \left(\ln(\theta_{\text{in}} \theta_{\text{out}}) + r_i s_j \ln\left(\frac{\theta_{\text{in}}}{\theta_{\text{out}}}\right) \right) \right. \\ &\quad \left. - \frac{d_i^{(X)} d_j^{(Y)}}{2} (\theta_{\text{in}} + \theta_{\text{out}} + r_i s_j (\theta_{\text{in}} - \theta_{\text{out}})) \right] \\ \implies \ell(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}) &= \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^l \left[A_{ij} \ln(\theta_{\text{in}} \theta_{\text{out}}) - d_i^{(X)} d_j^{(Y)} (\theta_{\text{in}} + \theta_{\text{out}}) \right. \\ &\quad \left. + \ln\left(\frac{\theta_{\text{in}}}{\theta_{\text{out}}}\right) \left(A_{ij} - d_i^{(X)} d_j^{(Y)} \cdot \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\ln \theta_{\text{in}} - \ln \theta_{\text{out}}} \right) s_i r_j \right]. \end{aligned}$$

We seek to maximise $\ell(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)})$ with respect to $G^{(X)}$ and $G^{(Y)}$ by choosing the co-community labels s_i and r_j . Therefore, we can drop the terms constant in s_i and r_j to give:

$$\tilde{\ell}(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}) = \sum_{i=1}^m \sum_{j=1}^l \left(A_{ij} - d_i^{(X)} d_j^{(Y)} \cdot \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\ln \theta_{\text{in}} - \ln \theta_{\text{out}}} \right) s_i r_j,$$

and defining:

$$\eta = \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\ln \theta_{\text{in}} - \ln \theta_{\text{out}}},$$

we therefore have:

$$\tilde{\ell}(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}) = \sum_{i=1}^m \sum_{j=1}^l \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) s_i r_j, \quad (45)$$

which we note as equivalent to equation 22 in [Newman, 2013]. Proceeding similarly to that work, by applying to equation 45 the constraints of equations 28 and 29 with Lagrange multipliers λ and μ and differentiating and equating to zero, we get:

$$\begin{aligned} \frac{\partial}{\partial s_i} \left[\sum_{i=1}^m \sum_{j=1}^l \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) s_i r_j - \lambda \sum_{i=1}^m d_i^{(X)} s_i^2 - \mu \sum_{j=1}^l d_j^{(Y)} r_j^2 \right] &= 0, \\ \frac{\partial}{\partial r_j} \left[\sum_{i=1}^m \sum_{j=1}^l \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) s_i r_j - \lambda \sum_{i=1}^m d_i^{(X)} s_i^2 - \mu \sum_{j=1}^l d_j^{(Y)} r_j^2 \right] &= 0, \end{aligned}$$

$$\begin{aligned} \implies \sum_{j=1}^l \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) r_j - 2\lambda d_i^{(X)} s_i &= 0, \\ \text{and } \sum_{i=1}^m \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) s_i - 2\mu d_j^{(Y)} r_j &= 0, \end{aligned}$$

$$\therefore \left(\mathbf{A} - \eta \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top \right) \mathbf{r} = 2\lambda \mathbf{D}^{(X)} \mathbf{s}, \quad (46)$$

$$\text{and } \left(\mathbf{A}^\top - \eta \mathbf{d}^{(Y)} \left(\mathbf{d}^{(X)} \right)^\top \right) \mathbf{s} = 2\mu \mathbf{D}^{(Y)} \mathbf{r}. \quad (47)$$

Combining equations 46 and 47 by substituting for s and r , and following simplification identical to equations 34 to 35, gives:

$$\begin{aligned} \mathbf{W}^\top \mathbf{W} \mathbf{r} &= 4\lambda \mu \mathbf{r}, \\ \text{and } \mathbf{W} \mathbf{W}^\top \mathbf{s} &= 4\lambda \mu \mathbf{s}, \end{aligned}$$

where

$$\mathbf{W} = \left(\mathbf{D}^{(X)} \right)^{-1/2} \left(\mathbf{A} - \eta \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top \right) \left(\mathbf{D}^{(Y)} \right)^{-1/2}.$$

Hence \mathbf{s} and \mathbf{r} are left and right singular vectors of the singular value decomposition of \mathbf{W} , again with corresponding singular values $4\lambda\mu$. Pre-multiplying 46 and 47 by $\mathbf{1} = (1, 1, 1, \dots)$ leads to:

$$\mathbf{r}^\top \mathbf{d}^{(Y)} (1 - d^{++}\eta) = 2\lambda \mathbf{s}^\top \mathbf{d}^{(X)}, \quad (48)$$

$$\text{and } \mathbf{s}^\top \mathbf{d}^{(X)} (1 - d^{++}\eta) = 2\lambda \mathbf{r}^\top \mathbf{d}^{(Y)}. \quad (49)$$

Substituting for $\mathbf{s}^\top \mathbf{d}^{(X)}$ and $\mathbf{r}^\top \mathbf{d}^{(Y)}$ in equations 49 and 48 gives:

$$\mathbf{s}^\top \mathbf{d}^{(X)} \left[(1 - d^{++}\eta)^2 - 4\mu\lambda \right] = 0,$$

$$\text{and } \mathbf{r}^\top \mathbf{d}^{(Y)} \left[(1 - d^{++}\eta)^2 - 4\mu\lambda \right] = 0,$$

and therefore because $(1 - d^{++}\eta)^2 - 4\mu\lambda$ is not guaranteed to be zero,

$$\mathbf{s}^\top \mathbf{d}^{(X)} = 0,$$

$$\text{and } \mathbf{r}^\top \mathbf{d}^{(Y)} = 0.$$

Therefore, equations 46 and 47 reduce to:

$$\begin{aligned} \mathbf{A}\mathbf{r} &= 2\lambda\mathbf{D}^{(x)}\mathbf{s} \\ \text{and } \mathbf{A}^\top\mathbf{s} &= 2\mu\mathbf{D}^{(y)}\mathbf{r}, \end{aligned}$$

and again combining these equations by substituting for \mathbf{s} and \mathbf{r} and following equivalent simplification to equations 34 to 35, we hence find that \mathbf{s} and \mathbf{r} are left and right singular vectors of the co-Laplacian (equation 7). Therefore, the choice of the co-community labels \mathbf{s} and \mathbf{r} which maximises the model likelihood specified in equation 3, subject also to the constraint of equation 2, is equivalent to the maximum co-modularity assignment obtained via Algorithm 1.

Supplement C: Proof of Lemma 1

Define \mathbf{A} , $k^{(X)}$, $k^{(Y)}$ according to Definition 1, define $\xi^{(X)}$ and $\xi^{(Y)}$ according to Definition 2, define f , \tilde{f} and γ according to Definition 5, and define ρ and \tilde{M} according to Lemma 1. Define bandwidths $h_p^{(X)} = |g_p^{(X)}|$ and $h_q^{(Y)} = |g_q^{(Y)}|$, where $|\cdot|$ represents cardinality, define $\omega(p, q)$ as the domain of integration over the block corresponding to the pairing of X -node grouping $g_p^{(X)}$ with Y -node grouping $g_q^{(Y)}$, and define $\bar{A}_{p,q}$ as the corresponding block average,

$$\bar{A}_{p,q} = \frac{\sum_{j \in g_q^{(Y)}} \sum_{i \in g_p^{(X)}} A_{ij}}{h_p^{(X)} \cdot h_q^{(Y)}}.$$

Then, the bias-variance decomposition of the MISE of the blockmodel approximation of the graphon function \hat{f} can be written as [Olhede and Wolfe, 2014]:

$$\begin{aligned} \text{MISE}(\hat{f}) &\leq \mathbb{E} \iint_{(0,1)^2} |f(x, y) - \hat{f}(x, y)|^2 dx dy = \\ &\sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \iint_{\omega(p,q)} \left\{ \left| f(x, y) - \frac{\mathbb{E}(\bar{A}_{p,q})}{\rho} \right|^2 + \frac{\text{Var}(\bar{A}_{p,q})}{\rho^2} \right\} dx dy. \quad (50) \end{aligned}$$

As well as specifying groupings of X and Y -nodes, $g_p^{(X)}$ and $g_q^{(Y)}$ imply mappings between the adjacency matrix margins and the graphon margins, and the domain of integration $\omega(p, q)$ is hence a contiguous region of the graphon, which maps to entries of the adjacency matrix which are not necessarily contiguous.

Modelling the equi-smooth graphon \tilde{f} as a linear stretch transformation of the anisotropic graphon f , by anisotropy factor γ , means that we can write:

$$f(x, y) = \tilde{f}(\gamma x, y/\gamma).$$

We define the graphon oracle [Wolfe and Olhede, 2013, Olhede and Wolfe, 2014] ordering of the X and Y -nodes according to $\xi^{(X)}$ and $\xi^{(Y)}$ respectively. These are unobservable latent random vectors, which map the locations of the X and Y nodes from the margins of the graphon to the margins of the adjacency matrix. I.e., $\xi_i^{(X)}$ and $\xi_j^{(Y)}$ provide the locations on the graphon margins which correspond to the X and Y -nodes i and j respectively, where i and j are the adjacency matrix indices of these nodes. We define $(i)^{-1}$ as a function which gives the rank of $\xi_i^{(X)}$, $1 \leq i \leq m$, and similarly $(j)^{-1}$ as a function which gives the rank of $\xi_j^{(Y)}$, $1 \leq j \leq l$. Therefore, $(i)^{-1}$ and $(j)^{-1}$ are functions which take the ordering along the adjacency matrix margins, and return the ordering along the graphon margins. Hence, the inverses of these functions, (i) and (j) , take the ordering along the graphon margins, and return the corresponding ordering along the adjacency matrix margins. Adapting the proof of Lemma 3 from [Olhede and Wolfe, 2014] to the anisotropic graphon, by defining $i_m = i/(m+1)$ and $j_l = j/(l+1)$, and assuming that \tilde{f} is Lipschitz-continuous, gives:

$$\begin{aligned} \left| f\left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}\right) - f(i_m, j_l) \right| &= \left| \tilde{f}\left(\gamma \xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}/\gamma\right) - \tilde{f}(\gamma i_m, j_l/\gamma) \right| \\ &\leq \tilde{M} \left| \left(\gamma \xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}/\gamma\right) - (\gamma i_m, j_l/\gamma) \right|. \end{aligned}$$

Writing the variances and applying Jensen's inequality as in [Olhede and Wolfe, 2014] we get,

$$\begin{aligned}
\text{Var} \left(\xi_{(i)}^{(X)} \right) &= \frac{i_m(1-i_m)}{m+2} \leq \frac{1/4}{m+2}, \\
\text{Var} \left(\xi_{(j)}^{(Y)} \right) &= \frac{j_l(1-j_l)}{l+2} \leq \frac{1/4}{l+2}, \\
\Rightarrow \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left\{ \gamma^2 \left(\xi_{(i)}^{(X)} - i_m \right)^2 + \frac{1}{\gamma^2} \left(\xi_{(j)}^{(Y)} - j_l \right)^2 \right\}^{\frac{1}{2}} \\
&\leq \left(\gamma^2 \text{Var} \left(\xi_{(i)}^{(X)} \right) + \frac{1}{\gamma^2} \text{Var} \left(\xi_{(j)}^{(Y)} \right) \right)^{\frac{1}{2}} \\
&\leq \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}}, \\
\therefore \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left| f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) - f(i_m, j_l) \right| &\leq \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}}. \quad (51)
\end{aligned}$$

Now adapting Lemma 2 from [Olhede and Wolfe, 2014], we apply the law of iterated expectations to $A_{(i)(j)}$, to obtain:

$$\mathbb{E} \left(A_{(i)(j)} \right) = \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\mathbb{E}_{A|\xi^{(X)}, \xi^{(Y)}} \left(A_{(i)(j)} \middle| \xi^{(X)}, \xi^{(Y)} \right) \right] = \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right], \quad (52)$$

then using Jensen's inequality we get:

$$\left| \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] - \rho f(i_m, j_l) \right| \leq \rho \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\left| f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) - f(i_m, j_l) \right| \right], \quad (53)$$

and hence combining equations 51-53, we have:

$$\left| \mathbb{E} \left(A_{(i)(j)} \right) - \rho f(i_m, j_l) \right| \leq \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}}. \quad (54)$$

Now applying the law of total variance to $A_{(i)(j)}$, as in Lemma 2 from [Olhede and Wolfe, 2014], we obtain:

$$\begin{aligned}
\text{Var} \left(A_{(i)(j)} \right) &= \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\text{Var}_{A|\xi^{(X)}, \xi^{(Y)}} \left(A_{(i)(j)} \middle| \xi^{(X)}, \xi^{(Y)} \right) \right] \\
&\quad + \text{Var}_{\xi^{(X)}, \xi^{(Y)}} \left[\mathbb{E}_{A|\xi^{(X)}, \xi^{(Y)}} \left(A_{(i)(j)} \middle| \xi^{(X)}, \xi^{(Y)} \right) \right] \\
&= \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \left(1 - \rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right) \right] \\
&\quad + \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho^2 \left(f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right)^2 \right] - \left(\mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \right)^2 \\
&= \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] - \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho^2 \left(f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right)^2 \right] \\
&\quad + \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho^2 \left(f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right)^2 \right] - \left(\mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \right)^2 \\
&= \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \left\{ \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[1 - \rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \right\}. \quad (55)
\end{aligned}$$

From equation 51, we get:

$$\mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \leq \rho f(i_m, j_l) + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \quad (56)$$

and

$$- \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[\rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \leq -\rho f(i_m, j_l) + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}}, \quad (57)$$

and hence also

$$\mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[1 - \rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \geq 1 - \rho f(i_m, j_l) - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \quad (58)$$

and

$$- \mathbb{E}_{\xi^{(X)}, \xi^{(Y)}} \left[1 - \rho f \left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)} \right) \right] \geq -1 + \rho f(i_m, j_l) - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}}. \quad (59)$$

Now combining equation 56 with the negative of equation 59 and applying equation 55 we get:

$$\begin{aligned} \text{Var}(A_{(i)(j)}) &\leq \left[\rho f(i_m, j_l) + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right] \\ &\quad \cdot \left[1 - \rho f(i_m, j_l) + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right] \end{aligned}$$

and hence:

$$\begin{aligned} \text{Var}(A_{(i)(j)}) &\leq \rho f(i_m, j_l) [1 - \rho f(i_m, j_l)] \\ &\quad + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \left[1 + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right]. \end{aligned} \quad (60)$$

Similarly combining the negative of equation 57 with equation 58 and applying equation 55 we get:

$$\begin{aligned} \text{Var}(A_{(i)(j)}) &\geq \left[\rho f(i_m, j_l) - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right] \\ &\quad \cdot \left[1 - \rho f(i_m, j_l) - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right], \end{aligned}$$

and hence:

$$\begin{aligned} \text{Var}(A_{(i)(j)}) &\geq \rho f(i_m, j_l) [1 - \rho f(i_m, j_l)] \\ &\quad - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \left[1 - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right], \end{aligned}$$

and therefore:

$$\begin{aligned} -\text{Var}(A_{(i)(j)}) &\leq -\rho f(i_m, j_l) [1 - \rho f(i_m, j_l)] \\ &\quad + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \left[1 - \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right] \\ &\leq -\rho f(i_m, j_l) [1 - \rho f(i_m, j_l)] \\ &\quad + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \left[1 + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right], \end{aligned} \quad (61)$$

and hence combining equations 60 and 61 we get:

$$\begin{aligned} &|\text{Var}(A_{(i)(j)}) - \rho f(i_m, j_l) [1 - \rho f(i_m, j_l)]| \\ &\leq \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \cdot \left[1 + \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}^{\frac{1}{2}} \right]. \end{aligned} \quad (62)$$

Now referring to equation 54 and comparing it to equation 6 of Supporting Information Section A in [Olhede and Wolfe, 2014], allows us to re-write the covariance expression in Lemma 2 of [Olhede and Wolfe, 2014] giving:

$$\text{Cov} (A_{(i)(j)}, A_{(i')(j')}) \leq \rho^2 \widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2} \cdot \frac{1}{4(l+2)} \right\}, \quad (63)$$

$i \neq i', j \neq j'$. We can then use equations 54, 62 and 63 to adapt Proposition 1 from [Olhede and Wolfe, 2014], denoting the average values of f and f^2 over the block corresponding to the pairing of $g_p^{(X)}$ with $g_q^{(Y)}$ as $\bar{f}_{p,q}$ and $\bar{f}_{p,q}^2$ respectively,

$$\bar{f}_{p,q} = \frac{1}{|\omega(p,q)|} \iint_{\omega(p,q)} f(x,y) dx dy \quad (64)$$

and

$$\bar{f}_{p,q}^2 = \frac{1}{|\omega(p,q)|} \iint_{\omega(p,q)} f^2(x,y) dx dy, \quad (65)$$

where

$$|\omega(p,q)| = \frac{h_p^{(X)}}{m} \cdot \frac{h_q^{(Y)}}{l},$$

to give:

$$|\mathbb{E}(\bar{A}_{p,q}) - \rho \bar{f}_{p,q}| \leq \rho \widetilde{M} \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}^{\frac{1}{2}} \{1 + o(1)\} \quad (66)$$

and

$$\begin{aligned} & \left| \text{Var}(\bar{A}_{p,q}) - \frac{\rho \bar{f}_{p,q} - \rho^2 \bar{f}_{p,q}^2}{h_p^{(X)} \cdot h_q^{(Y)}} \right| \\ & \leq \frac{\rho \widetilde{M}}{h_p^{(X)} \cdot h_q^{(Y)}} \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}^{\frac{1}{2}} \{1 + o(1)\} + \rho^2 \widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}, \end{aligned} \quad (67)$$

which is a conservative upper bound. Now substituting equation 67 back into equation 50, we get:

$$\begin{aligned} \text{MISE}(\hat{f}) & \leq \sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \iint_{\omega(p,q)} [|\{f(x,y) - \bar{f}_{p,q}\} \\ & \quad + \{\bar{f}_{p,q} - \mathbb{E}(\bar{A}_{p,q})/\rho\}|^2 + \frac{\bar{f}_{p,q} - \rho \bar{f}_{p,q}^2}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \\ & \quad + \frac{\widetilde{M}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}^{\frac{1}{2}} \{1 + o(1)\} + \widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}] dx dy, \end{aligned}$$

then substituting equation 66, integrating and rearranging, leads to:

$$\begin{aligned} \text{MISE}(\hat{f}) & \leq \sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \left[\iint_{\omega(p,q)} |f(x,y) - \bar{f}_{p,q}|^2 dx dy \right. \\ & \quad + \left(2 \widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\} \{1 + o(1)\} + \frac{\bar{f}_{p,q} - \rho \bar{f}_{p,q}^2}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \right. \\ & \quad \left. \left. + \frac{\widetilde{M}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}^{\frac{1}{2}} \{1 + o(1)\} \right) \cdot \frac{h_p^{(X)}}{m} \cdot \frac{h_q^{(Y)}}{l} \right]. \end{aligned} \quad (68)$$

Then, adapting the proof of Lemma 1 from [Olhede and Wolfe, 2014], we can write:

$$\begin{aligned} |\bar{f}_{p,q} - f(x, y)| &= \left| \frac{1}{|\omega(p, q)|} \iint_{\omega(p, q)} f(x', y') dx' dy' - f(x, y) \right| \\ &\leq \frac{1}{|\omega(p, q)|} \iint_{\omega(p, q)} \left| \tilde{f}(\gamma x', y'/\gamma) - \tilde{f}(\gamma x, y/\gamma) \right| dx' dy'. \end{aligned}$$

Assuming \tilde{f} is Lipschitz continuous, it therefore follows that:

$$\begin{aligned} |\bar{f}_{p,q} - f(x, y)| &\leq \frac{1}{|\omega(p, q)|} \iint_{\omega(p, q)} \tilde{M} |(\gamma x', y'/\gamma) - (\gamma x, y/\gamma)| dx' dy' \\ &\leq \frac{1}{|\omega(p, q)|} \iint_{\omega(p, q)} \tilde{M} \sqrt{\gamma^2 \cdot \frac{(h_p^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h_q^{(Y)})^2}{l^2}} dx' dy' \\ \implies |\bar{f}_{p,q} - f(x, y)| &\leq \tilde{M} \sqrt{\gamma^2 \cdot \frac{(h_p^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h_q^{(Y)})^2}{l^2}} \end{aligned}$$

and therefore

$$\frac{1}{|\omega(p, q)|} \iint_{\omega(p, q)} |\bar{f}_{p,q} - f(x, y)|^2 \leq \tilde{M}^2 \left\{ \gamma^2 \cdot \frac{(h_p^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h_q^{(Y)})^2}{l^2} \right\},$$

and hence summing over all the blocks corresponding to all pairings of X -node groupings $g^{(X)} \in G^{(X)}$ with Y -node groupings $g^{(Y)} \in G^{(Y)}$, and assuming $h^{(X)}$ and $h^{(Y)}$ are both constants, we get:

$$\sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \iint_{\omega(p, q)} |\bar{f}_{p,q} - f(x, y)|^2 \leq \tilde{M}^2 \left\{ \gamma^2 \cdot \frac{(h^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h^{(Y)})^2}{l^2} \right\}. \quad (69)$$

Recalling equation 64 and equation 10, i.e.,

$$\iint_{(0,1)^2} f(x, y) dx dy = 1,$$

and noting that:

$$\sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \frac{\bar{f}_{p,q} - \rho \bar{f}_{p,q}^2}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \leq \sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \frac{\bar{f}_{p,q}}{\rho \cdot h^{(X)} \cdot h^{(Y)}},$$

we can see that:

$$\begin{aligned} \sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \frac{\bar{f}_{p,q} - \rho \bar{f}_{p,q}^2}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} &\leq \sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \frac{m \cdot l}{\rho \cdot (h^{(X)})^2 \cdot (h^{(Y)})^2} \cdot \frac{h^{(X)}}{m} \cdot \frac{h^{(Y)}}{l} \cdot \bar{f}_{p,q} \\ &= \frac{m \cdot l}{\rho \cdot (h^{(X)})^2 \cdot (h^{(Y)})^2} \sum_{q=1}^{k^{(Y)}} \sum_{p=1}^{k^{(X)}} \iint_{\omega(p, q)} f(x, y) dx dy \\ &= \frac{m \cdot l}{\rho \cdot (h^{(X)})^2 \cdot (h^{(Y)})^2} \iint_{(0,1)^2} f(x, y) dx dy \\ &= \frac{m \cdot l}{\rho \cdot (h^{(X)})^2 \cdot (h^{(Y)})^2}. \end{aligned} \quad (70)$$

Now substituting 69 and 70 into 68, and rearranging, we get:

$$\begin{aligned} \text{MISE}(\hat{f}) &\leq \widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{(h^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h^{(Y)})^2}{l^2} \right\} \\ &\quad + 2\widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\} \{1 + o(1)\} \\ &\quad + \frac{1}{\rho \cdot h^{(X)} \cdot h^{(Y)}} + \frac{\widetilde{M}}{\rho \cdot h^{(X)} \cdot h^{(Y)}} \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\}^{\frac{1}{2}} \{1 + o(1)\} \end{aligned}$$

and hence:

$$\begin{aligned} \text{MISE}(\hat{f}) &\leq \widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{(h^{(X)})^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{(h^{(Y)})^2}{l^2} \right\} \\ &\quad + 2\widetilde{M}^2 \left\{ \gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l} \right\} \{1 + o(1)\} + \frac{1}{\rho \cdot h^{(X)} \cdot h^{(Y)}} \{1 + o(1)\}. \end{aligned}$$

Supplement D: Proof of Proposition 3

We wish to compare the null hypothesis:

$$H_0 : \quad \gamma = 1$$

against the alternative hypothesis:

$$H_1 : \quad \gamma \neq 1$$

where $\gamma^2 = M_Y/M_X$. In practice, we can only compare the ratio of our estimates of the gradients, which can be written as:

$$\frac{\hat{M}_Y}{\hat{M}_X} = \frac{\hat{b}_X \hat{p}_Y l}{\hat{p}_X \hat{b}_Y m} = \frac{(p_x + \epsilon_p^{(x)})(b_y + \epsilon_b^{(y)})}{(b_x + \epsilon_b^{(x)})(p_y + \epsilon_p^{(y)})} \cdot \frac{l}{m} = \frac{p_x b_y}{b_x p_y} \cdot \frac{(1 + \frac{\epsilon_p^{(x)}}{p_x})(1 + \frac{\epsilon_b^{(y)}}{b_y})}{(1 + \frac{\epsilon_b^{(x)}}{b_x})(1 + \frac{\epsilon_p^{(y)}}{p_y})} \cdot \frac{l}{m}.$$

Applying a first-order Maclaurin expansion, and neglecting products of errors, gives:

$$\hat{\gamma}^2 = \frac{\hat{M}_Y}{\hat{M}_X} = \frac{\hat{b}_X \hat{p}_Y}{\hat{p}_X \hat{b}_Y} \cdot \frac{m}{l} \approx \frac{p_x b_y}{b_x p_y} \cdot \left(1 + \frac{\epsilon_p^{(x)}}{p_x} + \frac{\epsilon_b^{(y)}}{b_y} + \frac{\epsilon_b^{(x)}}{b_x} + \frac{\epsilon_p^{(y)}}{p_y} \right) \cdot \frac{l}{m}$$

and hence, assuming b_x and b_y , p_x and p_y , b_x and p_y , p_x and b_y are independent, we can test against the null distribution

$$\hat{\gamma}^2 = \frac{\hat{b}_X \hat{p}_Y l}{\hat{p}_X \hat{b}_Y m} \sim \mathcal{N}(1, \hat{\tau}^2),$$

where

$$\begin{aligned} \hat{\tau}^2 &= \frac{\widehat{\text{Var}}(\hat{b}_X)}{\hat{b}_X} + \frac{\widehat{\text{Var}}(\hat{p}_Y)}{\hat{p}_Y} + \frac{\widehat{\text{Var}}(\hat{p}_X)}{\hat{p}_X} + \frac{\widehat{\text{Var}}(\hat{b}_Y)}{\hat{b}_Y} \\ &\quad + 2 \frac{\widehat{\text{Cov}}(\hat{b}_X, \hat{p}_X)}{\hat{b}_X \hat{p}_X} + 2 \frac{\widehat{\text{Cov}}(\hat{b}_Y, \hat{p}_Y)}{\hat{b}_Y \hat{p}_Y}, \quad (71) \end{aligned}$$

where \hat{b}_X , \hat{p}_X , \hat{b}_Y , \hat{p}_Y and their variances and covariances are estimated from the linear model fits as described.